Titre: Title:	Partage d'infrastructures et convergence fixe/mobile dans les réseaux 3GPP de prochaine génération
Auteur: Author:	Stéphane Ouellette
Date:	2012
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Ouellette, S. (2012). Partage d'infrastructures et convergence fixe/mobile dans les réseaux 3GPP de prochaine génération [Thèse de doctorat, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/950/

UNIVERSITÉ DE MONTRÉAL

PARTAGE D'INFRASTRUCTURES ET CONVERGENCE FIXE/MOBILE DANS LES RÉSEAUX 3GPP DE PROCHAINE GÉNÉRATION

STÉPHANE OUELLETTE
DÉPARTEMENT DE GÉNIE INFORMATIQUE ET GÉNIE LOGICIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(GÉNIE INFORMATIQUE)
SEPTEMBRE 2012

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

PARTAGE D'INFRASTRUCTURES ET CONVERGENCE FIXE/MOBILE DANS LES RÉSEAUX 3GPP DE PROCHAINE GÉNÉRATION

présentée par : OUELLETTE Stéphane
en vue de l'obtention du diplôme de : Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de :

Mme. BELLAÏCHE, Martine, Ph.D., présidente
M. PIERRE, Samuel, Ph.D., membre et directeur de recherche
M. QUINTERO, Alejandro, Doct., membre
M. KHENDEK, Ferhat, Ph.D., membre

À mon épouse Isabelle,
pour son soutien indéfectible.

À mes fils Frédéric et Martin,
pour tout le bonheur qu'ils m'apportent.

REMERCIEMENTS

J'aimerais tout d'abord remercier mon directeur de recherche, Samuel Pierre, pour m'avoir soutenu, conseillé et guidé tout au long de mon projet de recherche. Il m'a inculqué la rigueur scientifique dont je fais maintenant preuve.

Je voudrais aussi souligner la contribution très importante de Ericsson Research Canada, en particulier Laurent Marchand et Suresh Krishnan, pour leurs précieux conseils et leur soutien tout au long de mes travaux de recherche.

Enfin, j'aimerais remercier mes collègues et amis du *LAboratoire de recherche en Réseautique et Informatique Mobile* (LARIM) pour leur collaboration. J'ai passé en leur compagnie de nombreuses heures à discuter des travaux de recherche de tous et chacun. Ils ont ainsi suscité en moi l'intérêt pour la recherche, le tout dans une ambiance des plus chaleureuses.

RÉSUMÉ

Le déploiement de la technologie cellulaire de quatrième génération a débuté par quelques projets pilotes, notamment en Suède et en Norvège, dans la première moitié de 2010. Ces réseaux offrent dans un premier temps l'accès à Internet uniquement et comptent sur les réseaux de deuxième et troisième génération existants pour le support de la téléphonie et de la messagerie texte. Ce ne sera donc qu'avec l'avènement du IP Multimedia Subsystem (IMS) que tous les services seront supportés par la nouvelle architecture basée entièrement sur IP.

Les réseaux mobiles de quatrième génération promettent aux usagers des taux de transfert au-delà de 100 Mbits/s en amont, lorsque l'usager est immobile, et le support de la qualité de service [1] permettant d'offrir des garanties de débit, délai maximum, gigue maximale et d'un taux de perte de paquets borné supérieurement. Ces réseaux supporteront efficacement les applications utilisant la géolocalisation afin d'améliorer l'expérience de l'usager.

Les terminaux d'aujourd'hui offrent un éventail de technologies radio. En effet, en plus du modem cellulaire, les terminaux supportent souvent la technologie Bluetooth® qui est utilisée pour connecter entre autres les dispositifs mains-libres et les écouteurs. De plus, la majorité des téléphones cellulaires sont dotés d'un accès Wi-Fi® permettant à l'usager de transférer de grands volumes de données sans engorger le réseau cellulaire. Toutefois, cet accès n'est souvent réservé qu'au réseau résidentiel de l'usager ou à celui de son lieu de travail. Enfin, une relève verticale est presque toujours manuelle et entraîne pour le mobile un changement d'adresse IP, ce qui ultimement a pour conséquence une déconnexion des sessions en cours.

Depuis quelques années, une tendance se profile au sein de l'industrie qui est connue sous de nom de *convergence des réseaux fixes et mobiles.* [2] Cette tendance vise à plus ou moins long terme d'offrir l'accès Internet et la téléphonie à partir d'un seul terminal pouvant se connecter à un réseau d'accès local ou au réseau cellulaire. À ce jour, très peu d'opérateurs (e.g., NTT Docomo) offrent des terminaux ayant la possibilité de changer de point d'accès. Toutefois, le point d'accès doit appartenir à l'usager ou se situe à son lieu de travail.

Par ailleurs, on remarque un mouvement de convergence selon lequel différents réseaux utilisés pour les services d'urgence (tels que la police, les pompiers et ambulanciers) sont progressivement migrés (en raison de leurs coûts prohibitifs) vers un seul réseau offrant un très haut niveau de redondance et de fiabilité. Les services d'urgence démontrent des besoins en QoS similaires à ceux des particuliers sauf qu'ils nécessitent un accès prioritaire, ce qui peut entraîner la déconnexion d'un usager non-prioritaire lors d'une situation de congestion.

1. Le terme anglais Quality of Service (QoS) sera dorénavant utilisé.
2. Le terme anglais Fixed-Mobile Convergence (FMC) sera dorénavant utilisé.

En plus des services publics qui tentent de réduire leurs coûts d'exploitation en partageant l'accès aux réseaux commerciaux de communications, les opérateurs de ces réseaux sont aussi entrés dans une phase de réduction de coûts. Cette situation résulte du haut niveau de maturité maintenant atteint par l'industrie des communications mobiles. Par exemple, l'image de marque ou la couverture offerte par chacun d'eux ne constituent plus en soi un argument de vente suffisant pour attirer une nouvelle clientèle. Ceux-ci doivent donc se distinguer par une offre de services supérieure à celle de leur compétition.

Les opérateurs ont donc entrepris de sous-traiter des opérations non-critiques de leur entreprise afin de se concentrer sur l'aspect le plus profitable de cette dernière. Parallèlement à cette tendance, les opérateurs ont commencé à partager une portion de plus en plus importante de leurs infrastructures physiques avec leurs compétiteurs. Dans un premier temps, le partage s'est limité aux sites des stations de base et aux mâts qui supportent les antennes. Puis vint le partage des abris pour réduire les coûts de climatisation et d'hébergement des équipements. Ensuite, les opérateurs se mirent à partager les équipements radio, chacun contrôlant toutefois ses propres bandes de fréquences... Le partage des infrastructures physiques au-delà du premier nœud du réseau cœur n'est pas actuellement supporté en standardisation.

Il existe une autre tendance dans l'industrie des communications mobiles : le phénomène de spécialisation des opérateurs (i.e., la définition d'une clientèle cible pour ces derniers), ce qui entraîne des pics d'utilisation des ressources qui diffèrent selon la clientèle (e.g., gens d'affaires, jeunes et retraités). Ces opérateurs ont donc intérêt à partager certaines infrastructures physiques puisque le dimensionnement même des réseaux dépend largement de la charge maximale à traiter. Ceci a pour conséquence que la charge moyenne du réseau augmente sans que la charge maximale n'augmente significativement et permet de mieux rentabiliser les investissements colossaux qui sont requis pour le déploiement d'un réseau.

Les propositions existantes d'architectures de réseaux de prochaine génération [3] ont toutes comme point en commun d'être basées sur un réseau cœur tout-IP, d'offrir une QoS aux applications et une performance de l'ordre de 100 Mbits/s. De plus, ces dernières proposent des mécanismes de gestion des politiques [4] qui définissent l'utilisation des services offerts aux abonnés ainsi que la façon de comptabiliser l'usage des ressources du réseau.

On dénombre trois grandes catégories de politiques : celles se rattachant à l'usager (e.g., les abonnements or/argent/bronze, accès facturé *vs.* prépayé), celles qui dépendent du service demandé (e.g., pour un service donné, la bande passante maximale, la classe de service et la priorité d'allocation et de rétention des ressources) et enfin les politiques relatives à l'état du réseau (e.g., niveau de congestion, répartition des agrégats de trafic, etc).

3. Le terme anglais Next Generation Network (NGN) sera dorénavant utilisé.
4. Le terme anglais Policy and Charging Control (PCC) sera dorénavant utilisé.

Dans un premier article dont le titre est « *A Potential Evolution of the Policy and Charging Control/QoS Architecture for the 3GPP IETF-based Evolved Packet Core* », les aspects de FMC ainsi que du partage du réseau cœur sont traités conjointement puisqu'il faut que l'architecture PCC reflète les réalités des tendances de l'industrie décrites précédemment.

Suite à la description des tendances de l'industrie furent présentés les requis d'une architecture PCC qui rendent possibles la convergence des services (capacité d'utiliser un service à partir de n'importe quel accès), le partage du réseau cœur par plusieurs opérateurs mobiles virtuels [5], la création de politiques propres à chaque réseau d'accès ainsi que la micro-mobilité efficace des usagers dans les scénarios d'itinérance.

Dans un second temps, deux architectures de NGN furent évaluées en fonction des requis énumérés ci-dessus. Cette étude permit de déterminer qu'une solution hybride (avec les avantages de chacune mais sans leurs défauts respectifs) constituait une piste de solution prometteuse qui servit de base à notre proposition.

La solution proposée atteint son but par une meilleure répartition des rôles d'affaires ainsi que par l'introduction d'une entité centrale de contrôle nommée Network Policy Function (NPF) au sein du réseau de transport IP. En effet, les rôles d'affaires définis (fournisseurs d'accès, de réseau cœur et de services) permettent la création de domaines de politiques et administratifs distincts. Ces rôles deviennent nécessaires dans les cas de partage d'infrastructures. Dans le cas contraire, ils sont compatibles avec le modèle vertical actuel d'opérateur ; ce dernier joue alors tous les rôles.

Quant à l'introduction du NPF dans le réseau cœur, celui-ci permet de séparer la gestion des politiques régissant le réseau de transport IP des usagers, des services et des réseaux d'accès. De plus, il permet le partage du réseau cœur de façon à respecter les ententes de services liant ce dernier à chaque opérateur virtuel ainsi que les ententes de services liant le réseau cœur et le(s) réseau(x) d'accès.

Par ailleurs, le NPF permet d'ajouter au réseau cœur des services avancés à partager entre plusieurs opérateurs. Parmi ces services, on retrouve des fonctions de transcodage audio/vidéo, des caches de fichiers (e.g., pouvant servir à la distribution de films), d'antivirus grâce à l'inspection approfondie des paquets, etc. L'avantage d'introduire ces services au niveau transport est de permettre autant aux applications IMS qu'aux autres d'en bénéficier.

Le second article intitulé « *A Network Policy Function Node for a Potential Evolution of the 3GPP Evolved Packet Core* » constitue une extension du premier article qui décrit en détail les tendances de l'industrie, les architectures de gestion de politiques existantes et leurs caractéristiques, et enfin offrit un survol de la solution. En contrepartie, le second article aborde beaucoup plus en détail les impacts de la solution proposée sur l'architecture

5. Le terme anglais Mobile Virtual Network Operator (MVNO) sera dorénavant utilisé.

existante. En effet, une contribution significative de ce second article est de dresser la liste exhaustive de toutes les simplifications potentielles que permet la proposition d'architecture.

La contribution majeure du second article est que la solution proposée peut être déployée immédiatement avec un minimum d'impacts. Effectivement, une petite modification à l'architecture proposée dans le premier article, au niveau des interfaces du NPF, permit cette avancée. En conséquence, cette modification réconcilie les deux variantes actuelles d'architecture basées sur les protocoles GPRS Tunneling Protocol (GTP) et Proxy Mobile IPv6 (PMIPv6).

Le dernier apport important du second article est la démonstration du fonctionnement interne du NPF lorsque ce dernier contrôle un réseau de transport basé sur un mécanisme de tunnels tels que Multi-Protocol Label Switching (MPLS) ou encore Provider Backbone Bridge-Traffic Engineering (PBB-TE). Un processus d'ingénierie de trafic permet aux flux de trafic de contourner une zone de congestion, de mieux balancer la charge du réseau et d'assurer que les exigences en QoS sont toujours respectées.

Le troisième article intitulé « *A MultiAccess Resource ReSerVation Protocol (MARSVP) for the 3GPP Evolved Packet System* » traite de QoS dans les scénarios de FMC, plus particulièrement des applications qui ne sont pas supportées par le réseau. Par exemple, toutes les applications pair-à-pair qui représentent une portion infime du volume de trafic total attribué à ce type d'application ou celles qui sont naissantes et encore méconnues.

Les réseaux de deuxième et troisième générations ont été conçus de telle sorte que l'usager fournit au réseau les paramètres de QoS de l'application. Toutefois, le nombre de combinaisons des paramètres de QoS était très élevé et trop complexe à gérer. Il en résulta que pour la quatrième génération il fut décidé que dorénavant ce seraient les serveurs d'applications dans le réseau qui fourniraient ces paramètres de QoS. De même, un nombre restreint de classes de services fut défini, ce qui eut pour résultat de simplifier énormément la gestion de la QoS.

Lorsque sont considérés les concepts de FMC, il devient évident que le mécanisme décrit ci-dessus ne s'applique qu'aux accès 3GPP. En effet, chaque type d'accès définit ses propres mécanismes qui doivent souvent être contrôlés par le réseau et non par l'usager. De plus, certains accès ne disposent d'aucun canal de contrôle sur lequel circule les requêtes de QoS. De même, les protocoles existants de QoS sont souvent lourds et définis de bout-en-bout ; ils ne sont donc pas appropriés à l'utilisation qui est envisagée. En conséquence, la solution proposée consiste en un nouveau protocole multiaccès de réservation de ressources.

MARSVP utilise le canal de données que l'on retrouve sur tous les accès et confine les échanges de messages entre l'usager et le premier nœud IP. Les besoins en QoS sont définis en fonction des QoS Class Indicators (QCIs) ce qui rend MARSVP simple à utiliser. Suite à une requête de réservation de ressources acceptée par le réseau, ce dernier configure l'accès et retourne au terminal les informations requises à l'envoi paquets (aux couches 2 et 3).

ABSTRACT

Fourth generation cellular networks trials have begun in the first half of 2010, notably in Sweden and Norway. As a first step, these networks only offer Internet access and rely on existing second and third generation networks for providing telephony and text messaging. It's only after the deployment of the IP Multimedia Subsystem (IMS) that all services shall be supported on the new all-IP architecture.

Fourth generation mobile networks should enable end users to benefit from data throughputs of at least 100 Mbps on the downlink, when the user is stationary, and of Quality of Service (QoS) support that allows guarantees on throughput, maximum delay, maximum jitter and on the packet loss rate. These networks will efficiently support applications that rely on geolocation in order to improve the user's Quality of Experience (QoE).

Today's terminals can communicate using several radio technologies. Indeed, in addition to the cellular modem, terminals often support the Bluetooth® technology which is used for connecting handsfree devices and headsets. Moreover, most cell phones feature a Wi-Fi® interface that enables users to transfer huge volumes of data without congesting the cellular network. However, Wi-Fi connectivity is often restricted to the user's home network or his workplace. Finally, a vertical handover is nearly always done manually and forces the terminal to change its IP address, which ultimately disrupts all active data sessions.

A trend has emerged a few years ago among the mobile communications industry known as Fixed-Mobile Convergence (FMC). FMC is a trend aiming to provide Internet access and telephony on a single device capable of switching between local- and wide-area networks. At this time, very few operators (e.g., NTT Docomo) offer terminals capable of switching to another access automatically. However, the access point must belong to the user or be installed in his workplace.

At the same time, another kind of convergence has begun in which the dedicated networks for public safety (such as police, fire prevention and ambulances) are being progressively migrated (because of their high operational costs) toward a single highly reliable and redundant network. Indeed, these services exhibit QoS requirements that are similar to residential costumers' except they need a prioritized access, and that can terminate a non-priority user's session during congestion situations.

In addition to the public services that seek to reduce their operational costs by sharing commercial communications networks, the network operators have also entered a cost reduction phase. This situation is a result of the high degree of maturity that the mobile communications industry has reached. As an example, the branding or the coverage offered

by each of them isn't a sufficient sales argument anymore to enroll new subscribers. Operators must now distinguish themselves from their competition with a superior service offering.

Some operators have already started to outsource their less profitable business activities in order to concentrate on their key functions. As a complement to this trend, operators have begun to share an ever increasing portion of their physical infrastructures with their competitors. As a first step, infrastructure sharing was limited to the base station sites and antenna masts. Later, the shelters were shared to further reduce the cooling and hosting costs of the equipments. Then, operators started to share radio equipments but each of them operated on different frequency bands... Infrastructure sharing beyond the first core network node isn't actually supported in standardization.

There is an additional trend into the mobile communications industry which is the specialization of the operators (i.e., the identification of target customers by the operators). As a result, these operators experience disjoint traffic peaks because their customer bases have different behaviors. The former have a strong incentive to share infrastructures because network dimensioning mostly depends on the peak demand. Consequently, sharing infrastructures increases the average traffic load without significantly increasing the peak load because the peaks occur at different times. This allows operators to boost their return on investment.

Every existing Next Generation Network (NGN) architecture proposal features an all-IP core network, offers QoS to applications and a bandwidth on the downlink in the order of 100 Mbps. Moreover, these NGNs propose a number of Policy and Charging Control (PCC) mechanisms that determine how services are delivered to the subscribers and what charging method to apply.

There are three main categories of policies: those that are related to the subscriber (e.g., gold/silver/bronze subscription, prepaid *vs.* billed access), those that apply to services (e.g., for a given service, bandwidth limitation, QoS class assignment, allocation and retention priority of resources) and finally policies that depend on the current state of the network (e.g., congestion level, traffic engineering, etc).

In a first paper entitled "*A Potential Evolution of the Policy and Charging Control/QoS Architecture for the 3GPP IETF-based Evolved Packet Core*", FMC and Core Network (CN) sharing aspects are treated simultaneously because it is important that the logical PCC architecture reflects the realities of the industry trends described above.

Following the description of the trends in the communications industry were presented a list of four requirements that enable for a PCC architecture: service convergence (capacity to use a service from any type of access), CN sharing that allows several Mobile Virtual Network Operators (MVNOs) to coexist, the creation of local access network policies as well as efficient micro-mobility in roaming scenarios.

As a second step, two NGN architectures were evaluated upon the requirements mentionned above. This evaluation concluded that a hybrid solution (based on the key features of each architecture but without their respective drawbacks) would offer a very promising foundation for a complete solution.

The proposed solution achieved its goal with a clearer separation of the business roles (e.g., access and network providers) and the introduction of a Network Policy Function (NPF) for the management of the CN. Indeed, the business roles that were defined allow the creation of distinct policy/QoS and administrative domains. The roles become mandatory in infrastructure sharing scenarios. Otherwise, they maintain the compatibility with the actual vertically-integrated operator model; the latter then plays all of the business roles.

Introducing the NPF into the CN enables the CN policy management to be separated from policy management related to subscribers, services and access networks. Additionally, the NPF allows the CN to be shared by multiple Network Service Providers (NSPs) and respect the Service Level Agreements (SLAs) that link the IP Aggregation Network (IPAN) to the NSPs, as well as those that tie the IPAN to the Access Network Providers (ANPs).

Another benefit of the NPF is that it can share a number of advanced functions between several NSPs. Those functions include audio/video transcoding, file caches (e.g., that can be used for multimedia content delivery), Deep Packet Inspection (DPI) antivirus, etc. The main advantage to integrate those infrastructure services at the IP transport level is to allow both IMS and non-IMS applications to benefit from them.

A second paper entitled "*A Network Policy Function Node for a Potential Evolution of the 3GPP Evolved Packet Core*" constitutes an extension of the first paper that extensively described the industry trends, two existing PCC architectures and their characteristics, and finally offered an overview of the proposed solution. On the other hand, the second paper thoroughly describes all of the impacts that the proposal has on the existing 3GPP PCC architecture. Indeed, a significant contribution of this second paper is that it provides an extensive list of potential simplifications that the proposed solution allows.

The main contribution of the second paper is that from now on the proposed solution can be deployed over an existing PCC architecture with a minimum of impacts. Indeed, a small modification to the NPF's reference points enables this enhancement. As a consequence, this enhancement provided a solution that is compatible with both PCC architecture variants, based on either GPRS Tunneling Protocol (GTP) or Proxy Mobile IPv6 (PMIPv6).

A last contribution of the second paper is to demonstrate the NPF's internals when the former is controlling a an IPAN based on tunneling mechanisms such as Multi-Protocol Label Switching (MPLS) or Provider Backbone Bridge-Traffic Engineering (PBB-TE). A traffic engineering process allows traffic flow aggregates to pass around a congested node, to better

balance the load between the network elements and make sure that the QoS requirements are respected at all times.

The third paper entitled "*A MultiAccess Resource ReSerVation Protocol (MARSVP) for the 3GPP Evolved Packet System*" deals with QoS provisioning in FMC scenarios, especially for applications that are not directly supported by the network. As an example, all peer-to-peer applications (such as online gaming) that represent a small fraction of the total peer-to-peer traffic or those that are new and relatively unknown.

Second and third generation networks were designed such that the User Equipment (UE) would provide the network with the application's QoS parameters. However, the number of possible combinations of QoS parameters was very large and too complex to manage. As a result, for the fourth generation of networks, an application server would provide the PCC architecture with the right QoS parameters. In addition, a limited number of QoS classes were defined which in the end greatly simplified QoS management.

When FMC aspects are taken into account, it becomes trivial that the above mechanism only applies to 3GPP accesses. Indeed, each access type uses its own mechanisms that must often be controlled by the network instead of the user. Moreover, some accesses don't feature a control channel on which QoS reservation requests would be carried. Also, existing QoS protocols are often too heavy to support and apply end-to-end (between the sender and the receiver); they are not adequate for the intended use. As a consequence, the proposed solution is a new multiaccess resource reservation protocol.

The MultiAccess Resource ReSerVation Protocol (MARSVP) uses the data channel found on every access type and restricts the exchanges of signaling messages between the UE and the first IP hop. QoS requirements are expressed in terms of QoS Class Indicators (QCIs) which make MARSVP easy to use. Following a resource reservation that was accepted by the network, the latter configures the access point and informs the UE about the QoS informations required in order to send packets (at layers 2 and 3).

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

2G	2nd Generation
3G	3rd Generation
3GPP	3rd Generation Partnership Project (`http://www.3gpp.org`)
3GPP2	3rd Generation Generation Partnership Project 2 (`http://www.3gpp2.org`)
4G	4th Generation
AAA	Authentication, Authorization and Accounting
AC	Access Category
ADC	Application Detection and Control
AEG	Access Edge Gateway
AF	Application Function
AMBR	Aggregate Maximum Bit Rate
AN	Access Network
ANDSF	Access Network Discovery and Selection Function
ANP	Access Network Provider
AP	Access Point
API	Application Programming Interface
APN	Access Point Name
A-RACF	Access RACF
ARP	Allocation and Retention Priority
ATM	Asynchronous Transfer Mode
AuC	Authentication Center
AVP	Attribute-Value Pair
BBERF	Bearer Binding and Event Reporting Function
BBF	Bearer Binding Function
BGF	Border Gateway Function
BSC	Base Station Controller
BSS	Base Station Subsystem

BTS	Base Transceiver Station
CAPEX	Capital Expenditures
CBC	Cell Broadcast Center
CDMA	Code-Division Multiple Access
CDR	Charging Data Record
CLF	Connectivity session Location and repository Function
CFP	Contention-Free Period
CN	Core Network
CNG	Customer Network Gateway
CP	Contention Period
CPU	Central Processing Unit
CR-LDP	Constraint-Routing Label Distribution Protocol (RFC 5036) (RFC 3212)
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CTS	Clear To Send
C-RACF	Core RACF
DCF	Discrete Coordination Function
DHCP	Dynamic Host Configuration Protocol
DHCPv4	DHCP for IPv4 (RFC 2131)
DHCPv6	DHCP for IPv6 (RFC 3315)
DiffServ	Differentiated Services (RFC 2475)
DIFS	DCF InterFrame Space
DL	Downlink
DLS	Direct Link Setup
DPI	Deep Packet Inspection
DSCP	DiffServ Code Point (RFC 2474)
DSL	Digital Subscriber Line (`http://www.broadband-forum.org`)
DSMIPv6	Dual Stack Mobile IPv6 (RFC 5555)
ECMP	Equal Cost MultiPath
EDCA	Enhanced Distributed Channel Access
EDGE	Enhanced Data rates for GSM Evolution

EIR	Equipment Identity Register
eNodeB	LTE base station
ePDG	evolved Packet Data Gateway
EPC	Evolved Packet Core
EPS	Evolved Packet System
ETSI	European Telecommunications Standards Institute (`http://www.etsi.org`)
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FDD	Frequency-Division Duplexing
FDMA	Frequency-Division Multiple Access
FMC	Fixed-Mobile Convergence
FR	Frame Relay
FSC	Fiber Switch Capable (interface)
GBR	Guaranteed Bit Rate
GENI	Global Environment for Network Innovations (`http://www.geni.net`)
GERAN	GSM EDGE Radio Access Network
GGSN	Gateway GPRS Support Node
GMPLS	Generalized MPLS (RFC 3945)
GPRS	General Packet Radio Service
GRE	Generic Routing Encapsulation (RFC 2784)
GSM	Global System for Mobile communications (`http://www.gsmworld.com`)
GTP	GPRS Tunneling Protocol
GWCN	Gateway Core Network
HCCA	HCF Controlled Channel Access
HCF	Hybrid Coordination Function
HLR	Home Location Register
HRPD	High Rate Packet Data
HSS	Home Subscriber Server
HSPA	High Speed Packet Access
HSPA+	Enhanced HSPA

HTTP	HyperText Transfer Protocol (RFC 2616)
IaaS	Infrastructure as a Service
ICMP	Internet Control Message Protocol (RFC 792)
IEEE	Institute of Electrical and Electronics Engineers (`http://www.ieee.org`)
IETF	Internet Engineering Task Force (`http://www.ietf.org`)
IMS	IP Multimedia Subsystem
IntServ	Integrated Services (RFC 1633)
IP	Internet Protocol
IPv4	IP version 4 (RFC 791)
IPv6	IP version 6 (RFC 2460)
IPAN	IP Aggregation Network
IP-CAN	IP Connectivity Access Network
IP-MPLS	IP MPLS
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector (`http://www.itu.int/ITU-T`)
L2	Data-Link Layer
L2SC	Layer-2 Switch Capable (interface)
L3	Network Layer
LDP	Label Distribution Protocol (RFC 5036)
LER	Label Edge Router
LMA	Local Mobility Anchor
LMP	Link Management Protocol (RFC 4204)
LSC	Lambda Switch Capable (interface)
LSP	Label-Switched Path
LSR	Label-Switching Router
LTE	Long-Term Evolution
M2M	Machine-to-Machine (Communications)
MAC	Medium Access Control
MAG	Mobility Access Gateway
MARSVP	MultiAccess Resource ReSerVation Protocol

MBR	Maximum Bit Rate
MME	Mobility Management Entity
MNO	Mobile Network Operator
MOCN	Multi-Operator Core Network
MPLS	Multi-Protocol Label Switching (RFC 3031)
MPLS-TP	MPLS Transport Profile (RFC 5960)
MSTP	Multiple Spanning Tree Protocol
MVNO	Mobile Virtual Network Operator
NASS	Network Attachment SubSystem
NAT	Network Address Translation
NCP	Network Configuration Platform
NFC	Near Field Communication (`http://www.nfc-forum.org`)
NGN	Next Generation Network
NIST	National Institute of Standards and Technology
NPF	Network Policy Function
NSIS	Next Steps In Signaling (RFC 4080)
NSP	Network Service Provider
OAM	Operations, Administration and Maintenance
OCS	Online Charging System
OFCS	Offline Charging System
OPEX	Operational Expenditures
PBB	Provider Backbone Bridge (IEEE 802.1ah)
PBB-TE	Provider Backbone Bridge-Traffic Engineering (IEEE 802.1Qay)
PCC	Policy and Charging Control
PCE	Path Computation Element (RFC 4655)
PCEF	Policy and Charging Enforcement Function
PCF	Point Coordination Function
PCRF	Policy and Charging Rules Function
PDN	Packet Data Network
PDP	Packet Data Protocol

P2P	Peer-to-Peer
P-GW	Packet Data Network (PDN) Gateway
PHP	Penultimate Hop Popping
PIFS	PCF InterFrame Space
PMIPv6	Proxy Mobile IPv6 (RFC 5213)
PSC	Packet Switch Capable (interface)
QCI	QoS Class Indicator
QoE	Quality of Experience
QoS	Quality of Service
RACF	Resource and Admission Control Function
RACS	Resource and Admission Control Subsystem
RAN	Radio Access Network
RAT	Radio Access Type
RCEF	Resource Control Enforcement Function
RFID	Radio-Frequency Identification
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RSTP	Rapid Spanning Tree Protocol
RSVP	Resource reSerVation Protocol (RFC 2205)
RSVP-TE	Resource reSerVation Protocol (RFC 2205) for Traffic Engineering (RFC 3209)
RTCP	Real-Time Control Protocol (RFC 3550)
RTP	Real-Time Protocol (RFC 3550)
RTS	Request To Send
RTSP	Real-Time Streaming Protocol (RFC 2326)
SDF	Service Data Flow
S-GW	Serving Gateway
SFN	Single Frequency Network
SGSN	Serving GPRS Support Node
SIM	Subscriber Identity Module

SIP	Session Initiation Protocol (RFC 3261)
SLA	Service Level Agreement
SMTP	Simple Mail Transfer Protocol (RFC 5321)
SPDF	Service Policy Decision Function
SPR	Subscription Profile Repository
SRLG	Shared Risk Link Group
STP	Spanning Tree Protocol
TAI	Tracking Area Identity
TDD	Time-Division Duplexing
TDF	Traffic Detection Function
TDM	Time-Division Multiplex
TDMA	Time-Division Multiple Access
TE	Traffic Engineering
TED	Traffic Engineering Database
TISPAN	Telecommunications and Internet-converged Services and Protocols for Advanced Networks (`http://www.etsi.org/tispan`)
TXOP	Transmission Opportunity
UDC	User Data Convergence
UE	User Equipment
UDP	User Datagram Protocol (RFC 768)
UDR	User Data Repository
UMTS	Universal Mobile Telecommunications System (`http://www.gsmworld.com`)
UL	Uplink
UMB	Ultra Mobile Broadband
URL	Universal Resource Locator
UTRAN	Universal Terrestrial Radio Access Network
VLAN	Virtual Local Area Network (IEEE 802.1q)
VNC	Virtual Network Controller
VOD	Video On-Demand

W-CDMA Wideband Code-Division Multiple Access

WiMAX Worldwide Interoperability for Microwave Access (`http://www.wimaxforum.org`)

CHAPITRE 1

INTRODUCTION

Les réseaux mobiles de quatrième génération 4G permettront aux usagers d'accéder à une gamme de services évolués dont les besoins en bande passante sont très variés. De plus, leurs réseaux cœur seront basés sur le Internet Protocol (IP) qui servira de protocole commun à tous les réseaux d'accès. Cela aura pour avantage de réduire les dépenses d'investissements pour déployer de nouveaux services puisque ces derniers seront tous accessibles par IP.

1.1 Définitions et concepts de base

Cette section permet au lecteur d'aborder les concepts de base requis à la compréhension des éléments de cette thèse ainsi que de définir tous les termes essentiels. Suite à quelques définitions de termes couramment utilisés, nous expliquerons plus en détail les termes *Réseaux de prochaine génération*, *Convergence fixe/mobile* et *Qualité de service*.

Nomadisme : Capacité d'un usager à se connecter à son réseau d'attache à partir d'un emplacement extérieur. Un mécanisme associe son adresse fixe (associée à l'identité de l'usager et l'emplacement topologique habituel du terminal) à une adresse temporaire (associée à l'emplacement topologique actuel de l'usager). Le nomadisme est un cas particulier de mobilité dans lequel l'usager est immobile et ne peut effectuer de relève.

Mobilité : Capacité d'un usager à se déplacer alors qu'il est connecté à un réseau mobile. Un mécanisme de *relève* permet au terminal de changer de point d'accès (avec ou sans changement de technologie d'accès). Comme pour le nomadisme, une paire d'adresses (fixe et temporaire) est utilisée afin de joindre le mobile à tout moment.

Relève horizontale : Relève effectuée vers un point d'accès doté de la même technologie d'accès que le point d'accès actuel.

Relève verticale : Relève caractéristée par un changement de technologie d'accès.

Micro-mobilité : Mobilité restreinte à la couche 2 du réseau dans lequel se trouve le mobile. En conséquence, le terminal conserve son(ses) adresse(s) IP.

Macro-mobilité : Mobilité qui implique un changement d'adresses IP et de la couche 2.

Itinérance : État d'un terminal qui est connecté à son réseau d'attache via un réseau autre que le sien. Le terminal peut être fixe ou mobile.

1.1.1 Les réseaux de prochaine génération

L'organisme International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) utilise la définition suivante [35] : « *A Next Generation Network (NGN) is a packet-based network able to provide services including Telecommunication Services and able to make use of multiple broadband, Quality of Service (QoS)-enabled transport technologies and in which service-related functions are independent from underlying transport-related technologies. It offers unrestricted access by users to different service providers. It supports generalized mobility which will allow consistent and ubiquitous provision of services to users.* »

Les NGNs sont caractérisés par les aspects fondamentaux suivants :

- réseau de transport basé sur la commutation de paquets uniquement ;
- caractéristiques des services indépendantes de la technologie de transport sous-jacente ;
- l'usager accède à tous ses services à partir de n'importe quelle technologie d'accès ;
- débits comparables à ceux des technologies fixes avec QoS de bout-en-bout ;
- aucune restriction quant au choix du fournisseur de services par l'usager ;
- l'usager peut changer de point d'accès sans que ses sessions en cours ne soient affectées, et ce même si l'usager change de technologie d'accès au gré de ses déplacements ;
- l'accès aux services sera personnalisé en fonction du profil de l'usager ;
- conformité à toutes les réglementations, concernant par exemple les services d'urgence ou la confidentialité des communications.

1.1.2 Convergence Fixe/Mobile

La convergence des réseaux fixes et mobiles [1] est une tendance de l'industrie des communications mobiles qui se manifeste sous les quatre aspects décrits ci-dessous :

La *convergence des réseaux* signifie qu'il y a intégration des réseaux fixe et mobile, ainsi que de leurs services respectifs, dans le but de former les fondations d'un seul réseau de télécommunications. De plus, les réseaux de communications dédiés (e.g., pour les services d'urgence) ont de plus en plus tendance à utiliser les réseaux commerciaux en raison des coûts d'opération plus faibles. En conséquence, les réseaux se doivent d'être plus robustes et offrir un accès prioritaire aux unités d'urgence.

La *convergence des terminaux* implique que les terminaux intègrent de plus en plus les fonctionnalités qui étaient spécifiques à d'autres types d'appareils (e.g., téléphone, ordinateur et télévision) et que ces nouveaux terminaux sont dotés de plus d'une technologie d'accès.

1. Nous utiliserons désormais le terme anglais Fixed-Mobile Convergence (FMC).

La *convergence des services* est un concept qui implique que les services auxquels l'usager a souscrit peuvent être accédés à partir de n'importe quel type de technologie d'accès. De plus, ce concept sous-entend que les services sont définis au niveau IP et ne requièrent aucun mécanisme qui soit spécifique à un type d'accès pour leur bon fonctionnement. Enfin, ce type de convergence permet d'éviter la duplication inutile de certains services.

La *convergence commerciale* se manifeste par le regroupement du personnel affecté au marketing, à l'administration et au support technique des divisions fixe et mobile d'un opérateur. Ceci a pour but de provoquer un changement de mentalité des employés qui étaient auparavant assignés à l'une ou l'autre des divisions. En effet, la haute direction de cet opérateur doit amener ces employés à considérer les deux divisions comme étant complémentaires et non en compétition. Enfin, ce type de convergence peut aussi se manifester par l'alliance commerciale d'un opérateur fixe uniquement et d'un autre qui soit purement mobile.

1.1.3 La Qualité de Service

La Qualité de Service [2] comprend un ensemble de mécanismes qui permettent la différentiation de traitement appliqué aux paquets ou flux de données, et ce dans le but de satisfaire les besoins des applications. Les besoins en QoS d'une application peuvent être décrits en fonction des paramètres suivants :

- le **débit garanti** spécifie le débit minimum que doit fournir le réseau pour que l'application puisse s'exécuter correctement ;
- le **débit maximum** spécifie la limite supérieure de bande passante que l'application peut atteindre lors du fonctionnement normal de l'application ;
- le **délai maximal de bout-en-bout** représente la limite supérieure de d'intervalle de temps qui s'écoule entre le début de la transmission d'un paquet et la réception complète de celui-ci par le récepteur ;
- la **gigue maximale de bout-en-bout** (*jitter*) représente le plus grand écart au délai moyen de bout-en-bout que peut subir un paquet pour que l'application puisse fonctionner adéquatement [3] ;
- le **taux de perte de paquets** représente la fraction des paquets qui n'ont pu être reçus correctement au récepteur sur le nombre total de paquets émis (incluant les retransmissions effectuées par l'émetteur).

2. Nous utiliserons dorénavant le terme anglais Quality of Service (QoS).

3. Cette définition est la plus utilisée dans l'industrie. Notez qu'il n'y a aucun lien entre la gigue et la notion statistique de *variance*.

Le délai de bout-en-bout

Le délai de bout-en-bout correspond à la somme, pour chaque lien emprunté par un paquet, de quatre composantes décrites ci-dessous :

Le délai de transmission correspond au temps requis par un émetteur pour transmettre les bits d'un paquets, incluant tous les en-têtes des couches lien, réseau et transport. Ce délai est significatif pour les liens lents qui envoient des paquets de grande taille.

Le délai de propagation représente le temps requis pour qu'une information traverse un lien (incluant au besoin des répéteurs). Ce délai est important pour les liens intercontinentaux ou satellitaires (plus particulièrement pour les satellites géostationnaires). Il est déterminé par la longueur du lien ainsi que de la vitesse de propagation de l'information dans le médium.

Le délai de traitement correspond au temps requis pour qu'un nœud du réseau décide du traitement à appliquer au paquet. Ce temps est petit comparativement aux autres composantes et on peut le considérer à peu près constant pour chaque nœud.

Le délai d'attente représente le temps qu'un paquet en attente de transfert passe dans une file d'attente d'une interface physique. Ce délai dépend de la classe de service à laquelle appartient le paquet (il y a différentes priorités pour les files d'attente) et du nombre de paquets déjà en attente dans la file de transmission. Pour les cas relativement simples, il est possible de modéliser ce délai grâce à la théorie des files d'attente.

Paradigmes de QoS dans les réseaux IP

Parmi les standards de l'Internet Engineering Task Force (IETF), deux achitectures de QoS (et leurs paradigmes respectifs) sont proposées :

- Integrated Services (IntServ) qui est basée sur les *réservations de ressources* ;
- Differentiated Services (DiffServ) qui est basée sur la *différentiation de services*.

L'architecture IntServ est basée sur les réservations de ressources au niveau des routeurs. Un protocole de réservation de ressources est utilisé pour signaler les besoins en QoS de l'émetteur. Les messages de signalisation sont traités par les routeurs qui vérifient s'ils ont des ressources disponibles. L'avantage principal de ce modèle de QoS est qu'une réservation de ressources, une fois acceptée, peut être honorée quelle que soit la charge du routeur. L'inconvénient majeur de ce modèle est l'évolutivité puisqu'un routeur doit maintenir les états des réservations qui furent acceptées.

L'architecture DiffServ est basée sur la différentiation de services appliquée individuellement à chaque paquet. L'avantage majeur de ce modèle est qu'il ne nécessite aucun protocole de signalisation puisque l'information requise au traitement du paquet se trouve dans l'en-tête de ce dernier. En contrepartie, il faut implémenter un mécanisme strict de contrôle d'admission aux frontières du réseau afin de ne pas accepter plus de trafic que le réseau peut en traiter. En effet, dans un agrégat de trafic, un flot qui consomme beaucoup plus que sa juste part de bande passante peut impacter négativement la QoS de tous les autres flots.

1.1.4 Les réseaux cellulaires 3GPP

Le consortium 3rd Generation Partnership Project (3GPP) regroupe un ensemble de manufacturiers et d'opérateurs à travers le monde. [4] 3GPP produit les spécifications de réseaux mobiles de deuxième (2G), troisième (3G) et quatrième générations connus respectivements sous les noms de Global System for Mobile communications (GSM), Universal Mobile Telecommunications System (UMTS) et Evolved Packet System (EPS).

Évolution des technologies 3GPP

GSM est une technologie 2G dont l'interface radio est basée sur le multiplexage dans le temps (TDMA) ou en fréquence (FDMA). Spécifié dans les versions 96 à 98 de 3GPP. Permet initialement de transporter 8 kbits/s de données. L'ajout de General Packet Radio Service (GPRS) [5] permet une bande passante jusqu'à 144 kbits/s puis de Enhanced Data rates for GSM Evolution (EDGE) qui permet d'atteindre un maximum de 236 kbits/s.

UMTS est une technologie 3G dont l'interface radio est basée sur le multiplexage de codes (CDMA). Spécifié dans les versions 99 et 4 à 7 de 3GPP. Permet initialement de transporter 384 kbits/s de données. Les ajouts de High Speed Packet Access (HSPA) puis de Enhanced HSPA (HSPA+) permettent d'atteindre jusqu'à 42 Mbits/s.

EPS est le nom donné à l'architecture 3GPP pour les versions 8 et suivantes. Le Evolved Packet Core (EPC) est une évolution de GPRS et constitue un réseau cœur entièrement basé sur IP. [6] Enfin, une nouvelle technologie radio nommée Long-Term Evolution (LTE) s'ajoute aux accès existants GSM et Wideband Code-Division Multiple Access (W-CDMA). Le EPS correspond à la combinaison du EPC et du nouveau réseau d'accès radio basé sur LTE connu sous le nom de Evolved Universal Terrestrial Radio Access Network (E-UTRAN).

4. Les réseaux opérant selon les spécifications 3GPP comptent en 2011 près de 5 milliards d'abonnés.
5. GPRS est le nom donné à l'architecture à commutation de paquets des réseaux GSM et UMTS.
6. L'architecture à commutation de circuits a été complètement éliminée de sorte que tous les services sont maintenant offerts sur IP.

L'architecture de gestion des politiques

En plus de HSPA+, la version 7 a introduit une architecture de gestion des politiques afin de permettre un contrôle fin sur tous les aspects d'accès et de comptabilisation de l'utilisation des ressources par les usagers. L'architecture de Policy and Charging Control (PCC) est en grande partie supportée par le Policy and Charging Rules Function (PCRF) qui est un nœud logique qui agit à titre de serveur de politiques dynamiques.

Pour EPS, il existe deux variantes de l'architecture PCC (elles seront présentées au chapitre 2), selon le choix du protocole de tunnelisation qui transporte les paquets des usagers :

- la première version utilise le GPRS Tunneling Protocol (GTP) pour gérer la mobilité, la QoS ainsi que les différentes procédures de gestion des *bearers*[7] ;
- la seconde utilise Proxy Mobile IPv6 (PMIPv6) pour gérer la mobilité seulement.

1.2 Éléments de la problématique

Lors des premiers déploiements de réseaux cellulaires, la zone de couverture de chacun des opérateurs constituait un facteur de différentiation entre ces derniers. Aujourd'hui, ce sont plutôt les forfaits de services offerts par ceux-ci qui déterminent le choix d'un opérateur. De plus, la croissance de l'utilisation des ressources se fait sentir depuis que les téléphones intelligents ont envahi les marchés. De surcroît, le revenu moyen par abonné est demeuré approximativement le même depuis plusieurs années alors que les opérateurs font face à des coûts d'investissements additionnels.

Dans cette thèse, nous évoluerons à l'intérieur d'un cadre fixé par les grandes tendances de l'industrie de même que des technologies d'accès actuellement disponibles. Parmi les tendances de l'industrie, nous aborderons :

- le morcellement du modèle traditionnel à intégration verticale d'opérateur mobile ;
- spécialisation des opérateurs ;
- la convergence fixe/mobile.

Une tendance de fond a émergé au sein de l'industrie des communications mobiles qui découle directement de son haut degré de maturité. En effet, les opérateurs sont encouragés à se concentrer sur ce qu'ils considèrent être leurs points les plus forts et délaisser les opérations les moins rentables (ou celles qui ne permettent pas de les distinguer de leur compétiteurs). La sous-traitance est maintenant considérée par certains opérateurs comme un excellent moyen de réduire leurs coûts d'exploitation.

Il existe de nos jours des opérateurs mobiles virtuels[8] qui ne disposent d'aucune infra-

7. Un *bearer* est un tunnel IP logique liant le terminal et le réseau, doté de paramètres spécifiques de QoS.
8. Nous utiliserons dorénavant le terme anglais Mobile Virtual Network Operator (MVNO).

structure. En effet, ces opérateurs achètent des volumes de bande passante à l'un des grands opérateurs nationaux et la revendent à leurs clients. Ainsi, les grands opérateurs peuvent sécuriser leurs revenus tout en minimisant leurs coûts administratifs (e.g., puisqu'ils n'ont pas à offrir de support à la clientèle ni gérer la facturation). En contrepartie, les opérateurs virtuels peuvent espérer tirer leur épingle du jeu en n'ayant pas à supporter les coûts prohibitifs du déploiement d'un réseau complet.

Une autre tendance de fond a émergé dans l'industrie : la *spécialisation des opérateurs*. Cette tendance se manifeste par la création d'opérateurs qui visent une clientèle plus ciblée grâce à une offre de services qui lui est adaptée. Ces opérateurs sont tous des MVNOs car le fait de viser une clientèle restreinte est incompatible avec les investissements colossaux nécessaires pour déployer un réseau complet.

Dans un autre ordre d'idée, les terminaux modernes sont presque tous dotés de plus d'une technologie radio. En effet, la grande majorité de tous les terminaux vendus aujourd'hui supportent Bluetooth et Wi-Fi. Bluetooth est utilisé pour connecter un appareil mains-libres au terminal ainsi que pour les petits transferts de fichiers entre appareils. Wi-Fi est utilisé pour les grands transferts de données, l'accès à Internet ainsi que pour la voix sur IP. Toutefois, les accès Wi-Fi actuels ne supportent que très rarement les extensions permettant la QoS (IEEE 802.11e). En conséquence, un point d'accès Wi-Fi partagé ne peut être utilisé de nos jours que pour transporter des données qui ne requièrent aucune garantie de QoS.

Lorsque les terminaux seront tous capables de supporter la QoS sur Wi-Fi, il sera possible de décharger le réseau cellulaire grâce à des points d'accès Wi-Fi installés dans les résidences, là où la couverture cellulaire est déficiente ou encore dans les endroits où une grande concentration d'usagers doit être servie. Le déploiement à grande échelle de la FMC permettra ainsi à un opérateur d'offrir à leurs usagers soit la connexion la plus rapide ou celle dont le rapport qualité/prix est le meilleur.

Enfin, avec l'avènement du EPS, 3GPP a radicalement changé de mécanisme pour qu'un terminal obtienne de la QoS pour ses flux de données. En effet, jusqu'à la version 7, le terminal était le seul responsable de demander au réseau la création de *bearers* dédiés [9] avec une QoS spécifique. Depuis la version 8, un serveur d'application demande au réseau de créer un *bearer* dédié sans l'intervention du terminal. [10]

Puisque chaque technologie d'accès possède ses propres mécanismes de QoS et que ces mécanismes seront vraisemblablement contrôlés à partir du réseau et non du terminal, il faut informer le terminal des associations de classes de services 3GPP à celles du réseau d'accès sous-jacent. Par exemple, WiMAX utilise le concept de connexion et la station de base gère

9. Ce mécanisme est connu sous de nom de *Secondary Packet Data Protocol (PDP) context creation*.

10. Notez qu'il existe encore une procédure permettant à un mobile de créer des *bearers* dédiés mais 3GPP encourage fortement le recours à des serveurs d'applications.

l'allocation des tranches de temps selon les classes de services et la charge du réseau. En contrepartie, Wi-Fi utilise un mécanisme d'accès au canal radio sans connexion et basé sur la compétition.

Lorsque la QoS est supportée avec Wi-Fi, deux mécanismes sont proposés :

- Enhanced Distributed Channel Access (EDCA) qui est basé sur la différentiation de services (les stations ayant du trafic important ont un accès prioritaire au canal) ;
- HCF Controlled Channel Access (HCCA) pour lequel le point d'accès contrôle l'accès au canal au besoin afin de donner la chance aux stations enregistrées à ce service d'envoyer et recevoir des données avec garanties strictes de QoS.

Dans un autre ordre d'idée, si on retire complètement au mobile la capacité d'effectuer une réservation de ressources sans passer par un serveur d'application, alors toute une classe d'applications se verra dans l'impossibilité d'obtenir de la QoS pour son bon fonctionnement. En effet, les applications Peer-to-Peer (P2P) ne sont pas directement supportées par le réseau puisque ces applications ont recours à un serveur qui maintient une liste d'addresses des pairs actuellement accessibles mais les échanges de données se font directement entre ces derniers.

La version 8 de l'architecture 3GPP a introduit une nouvelle approche à la QoS : le recours à des QoS Class Indicators (QCIs). Ces classes de services sont définies de telle sorte qu'on peut les associer à des profils d'applications. Toutefois, étant donné que l'architecture 3GPP n'est définie qu'au niveau IP, l'association entre les QCIs et les classes de services aux niveaux 1 et 2 n'est pas définie, surtout dans un scénario d'itinérance où un opérateur autre que le sien pourrait avoir défini des associations différentes. En conséquence, d'après les faits mentionnés précédemment, il devient évident que l'architecture du réseau cœur se doive de refléter ces réalités.

Les travaux de recherche porteront sur l'architecture PCC de 3GPP pour les versions 8 et suivantes. Même si la majorité de nos propositions peuvent être applicables à la variante basée sur GTP, nous concentrerons nos efforts sur la variante PMIPv6 qui fut développée dans le but de permettre au EPC d'interfacer des accès non-3GPP. De nombreux problèmes pour lesquels il n'existe aucune solution actuellement ou seulement des solutions partielles ne seront pas considérés dans cette thèse.

Pour ce qui est de la FMC, les sujets suivants seront considérés hors sujet :

- la découverte de points d'accès qui permettent au terminal de se connecter au EPC avec la QoS désirée ;
- le routage lorsque le terminal est connecté à plus d'un accès simultanément ;
- l'authentification des usagers auprès du EPC via les accès non-3GPP ;
- la décision de procéder à une relève totale ou partielle.

Enfin, pour le partage d'infrastructures, les sujets suivants seront considérés hors sujet :
- l'implémentation de mécanismes de gestion de politiques pour les accès non-3GPP ;
- les mécanismes de (re)négociation dynamique des ententes de services entre les opérateurs et fournisseurs d'infrastructures.

1.3 Objectifs de recherche

La présente thèse a pour objectif principal de proposer des améliorations à l'actuelle architecture PCC pour les réseaux 3GPP. Plus spécifiquement, cette thèse vise à :

1. dresser une liste des tendances au sein de l'industrie des communications mobiles ;
2. proposer des améliorations à l'architecture PCC afin de simplifier le partage d'infrastructures du réseau cœur entre plusieurs opérateurs ;
3. proposer un mécanisme permettant de spécifier les besoins en QoS d'une application qui n'est pas directement supportée par le réseau de façon consistante, quel que soit l'accès utilisé ou le nombre de relèves inter-technologies.

Ces améliorations doivent permettre de mieux supporter la FMC de même que de supporter des scénarios élaborés de partage d'infrastructures du réseau cœur entre plusieurs opérateurs. De plus, ces modifications doivent, dans la mesure du possible, simplifier l'architecture PCC afin d'en maximiser le degré d'acceptabilité par l'industrie. Enfin, les propositions doivent permettre à l'architecture PCC de satisfaire les besoins actuels et futurs des opérateurs.

1.4 Plan de la thèse

Dans le présent chapitre, nous avons survolé les définitions et les concepts de base relatifs au sujet de recherche de cette thèse. Nous y avons aussi présenté les éléments de la problématique et fixé les objectifs de recherche qui devront être atteints. Les étapes suivantes sont décrites ci-dessous.

Dans un premier temps, un survol de l'architecture 3GPP sera effectué, incluant une description sommaire des nœuds de l'architecture PCC. Une liste des principaux mécanismes de contrôle de la QoS sera ensuite dressée pour les accès radio les plus connus. Les tendances de fond de l'industrie des communications mobiles seront décrites en détail puisque ce sont ces dernières qui justifient les améliorations proposées à l'architecture PCC. Enfin, nous présenterons de articles scientifiques ayant pour sujet le partage d'infrastructures ou la gestion de la QoS dans un contexte de FMC.

Dans le chapitre 3, nous décrirons les démarches de l'ensemble du travail de recherche. De plus, nous établirons des liens entre les articles et les objectifs de recherche.

Le chapitre 4 reproduit le texte intégral d'un article publié qui traite de partage des infrastructures du réseau cœur et de FMC dans le EPS. Nous effectuerons un survol des tendances de l'industrie des communications mobiles, analyserons les architectures PCC de 3GPP ainsi que le Resource and Admission Control Subsystem (RACS) de l'architecture de l'organisme Telecommunications and Internet-converged Services and Protocols for Advanced Networks (TISPAN). Ensuite, nous élaborerons une liste de requis qui devront être satisfaits pour permettre un meilleur support du partage des infrastructures ainsi que de la FMC. Par la suite, nous présenterons une proposition d'architecture hybride ainsi qu'un exemple détaillé qui permet de démontrer les avantages de notre solution.

Le chapitre 5 présente le texte intégral d'un second article qui améliore sensiblement l'architecture présentée dans le premier article. On y décrit en détail les impacts de notre solution sur chacun des nœuds du réseau cœur de l'architecture PCC existante. Plusieurs simplifications potentielles seront abordées.

Le chapitre 6 contient le texte intégral d'un troisième article qui porte sur la signalisation la de QoS initiée par le terminal. En effet, nous verrons pourquoi la nouvelle façon de signaler les besoins en QoS dans le EPS ne peut combler tous les besoins et que parfois seul le terminal peut disposer des informations de QoS requises au bon fonctionnement d'une application.

Suite à la présentation des trois articles, nous procéderons à une discussion générale en regard des résultats obtenus au cours de cette recherche et en lien avec la revue de littérature. Plus particulièrement, nous comparerons notre solution au partage d'infrastructures du réseau cœur avec celle d'une publication parue cinq mois après notre premier article.

Enfin, le chapitre 8 fera un survol des contributions de cette thèse avant d'aborder un ensemble de sujets à considérer lors de travaux futurs. En effet, nous proposerons des avenues de recherches ayant des liens directs avec notre proposition d'architecture. Par exemple, les *Infrastructure Services* que nous avons introduits dans notre architecture ne disposent d'aucun mécanisme standardisé pour découvrir ou invoquer ces services en question. Par ailleurs, nous énoncerons quelques réflexions concernant l'évolution future de l'industrie des communications mobiles et ses impacts sur nos travaux de recherche.

CHAPITRE 2

ARCHITECTURES, TECHNOLOGIES D'ACCÈS ET MÉCANISMES DE VIRTUALISATION LIÉS AU PARTAGE D'INFRASTRUCTURES

La première partie de ce chapitre décrira l'architecture générale des réseaux 3GPP. Ensuite, nous porterons notre attention sur l'architecture PCC qui est responsable de la gestion des ressources et de la comptabilisation de l'utilisation de ces dernières. Puis, nous présenterons un ensemble d'articles retrouvés dans la littérature scientifique qui abordent le partage d'infrastructures ainsi que la FMC.

2.1 Architecture générale du Evolved Packet System (EPS)

Le EPS est composé de deux sous-réseaux dont nous présenterons les détails ci-après :

1. le Radio Access Network (RAN) ;
2. le Evolved Packet Core (EPC).

Le reste de la présente section décrira brièvement chacun des nœuds. Pour les nœuds 2G et 3G, nous ne ferons qu'un rapide survol des fonctions de ceux-ci. De même, nous n'accorderons qu'une attention limitée aux infrastructures à commutation de circuits et ce dans le but de nous concentrer sur le GPRS. La Figure 2.1 présente l'architecture globale du EPS. Les éléments en pointillé ne peuvent être présents que lorsque l'interface S5 est basée sur PMIPv6. En effet, une première variante [5] découle directement d'une évolution de l'architecture de la version 7 et utilise le GTP au niveau de l'interface S5. Cette première variante ne supporte que les accès 3GPP. En contrepartie, une seconde variante [6] fut proposée afin d'interfacer les accès non-3GPP au EPC. Cette dernière est basée sur PMIPv6. Enfin, plusieurs sous-systèmes non-essentiels ne seront pas abordés.

2.1.1 Composantes du RAN

Le RAN comprend tous les éléments spécifiques aux technologies d'accès 3GPP. Les noms suivants réfèrent respectivement aux RANs 2G, 3G et 4G :

- GSM EDGE Radio Access Network (GERAN) ;
- Universal Terrestrial Radio Access Network (UTRAN) ;
- Evolved Universal Terrestrial Radio Access Network (E-UTRAN).

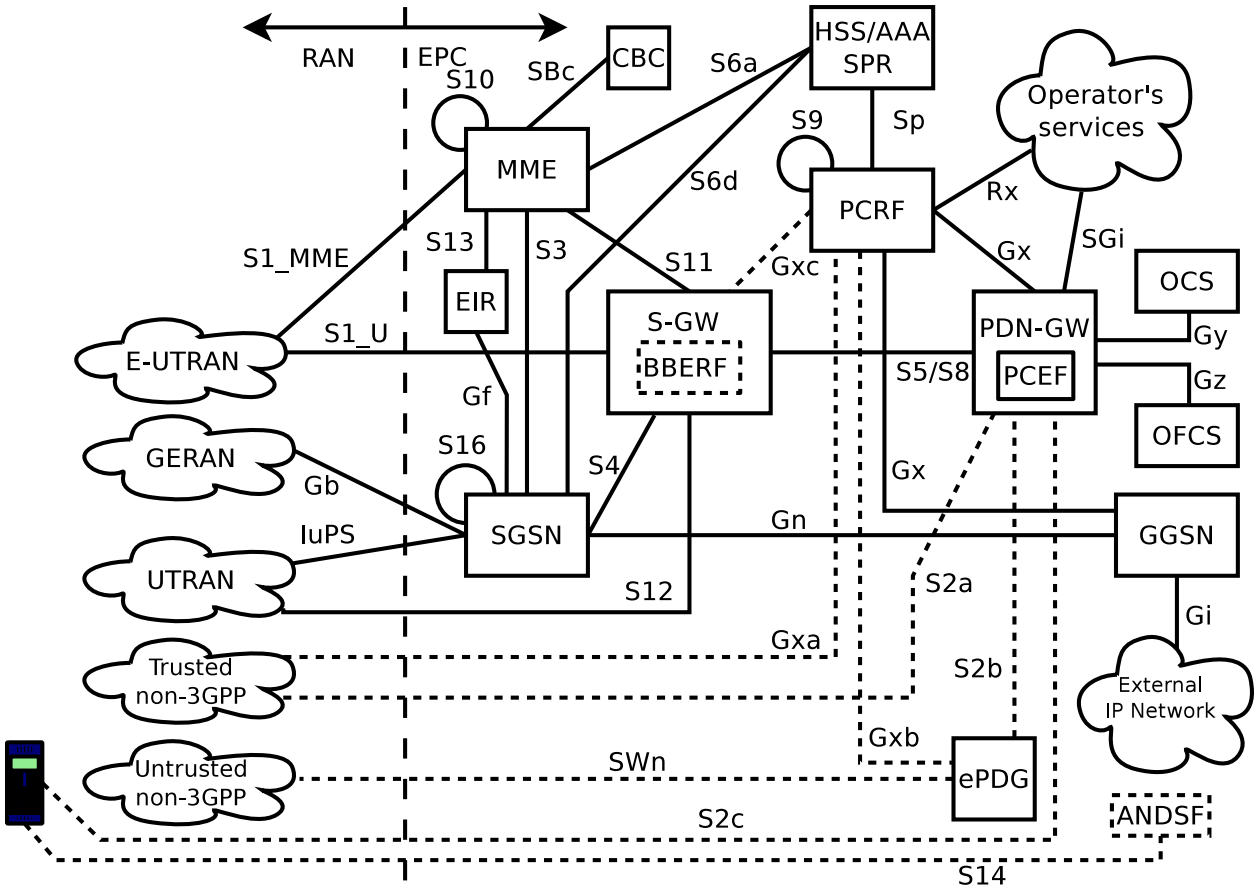


Figure 2.1 Architecture générale du Evolved Packet System (EPS)

Composantes du GERAN

Le GERAN comprend au moins un Base Station Subsystem (BSS). Chaque BSS contient plusieurs Base Transceiver Stations (BTSs) qui sont contrôlés par un seul Base Station Controller (BSC). Les BTSs contiennent tous les équipements d'interface radio (avec la possibilité de contrôler plusieurs fréquences dans le cas des stations sectorisées), les antennes, de même que les équipements nécessaires à l'encryption des données échangées avec le BSC. La Figure 2.2 illustre un BSS qui est connecté au Serving GPRS Support Node (SGSN) situé dans le réseau GPRS.[1] Cependant, la portion de l'architecture à commutation de circuits n'est pas présentée.

Le BSC peut contrôler plusieurs dizaines (et même au-delà d'une centaine de BTSs). Son rôle est d'allouer les canaux radio, prendre des mesures de qualité du signal reçu par le mobile et d'effectuer les relèves d'un BTS à un autre lorsque ceux-ci sont contrôlés par le même BSC.

1. Le SGSN sera présenté dans la sous-section 2.1.2.

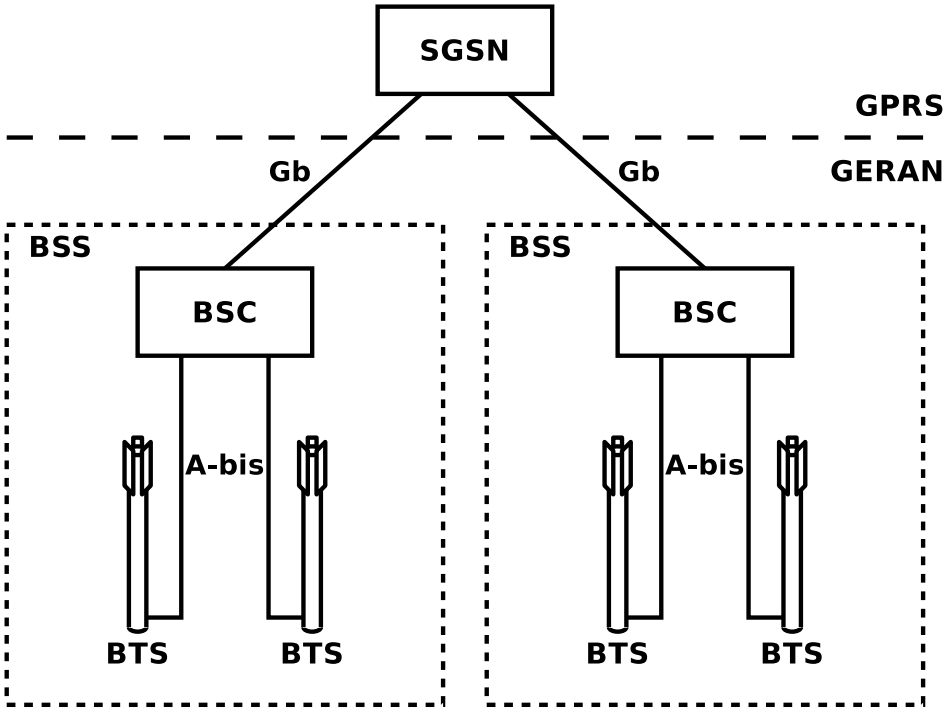


Figure 2.2 Architecture à commutation de paquets du GERAN

Composantes du UTRAN

Le UTRAN (voir la Figure 2.3) possède une architecture presqu'identique à celle du GERAN. En effet, le UTRAN est composé d'au moins un Radio Network Subsystem (RNS). Chaque RNS contient un Radio Network Controller (RNC) qui gère plusieurs Node Bs. Les Node Bs effectuent les mêmes tâches que les BTSs dans le GERAN. Toutefois, l'avènement de HSPA a fait en sorte que certaines fonctions (e.g., la retransmission de fragments de paquets en cas d'erreur) ont été intégrées dans le Node B afin de diminuer le temps de réponse.

Le RNC est l'équivalent d'un BSC. Il implémente quelques fonctions supplémentaires permettant par exemple la relève *soft* inter-RNC.[2] Aussi, depuis la version 7 de 3GPP, le RNC peut supporter le *Direct Tunnel Optimization* qui permet aux transferts de haut débits de passer outre le nœud SGSN pour échanger des données directement avec le Gateway GPRS Support Node (GGSN). Le SGSN n'est alors utilisé que pour la portion contrôle et peut ainsi évoluer de façon indépendante de l'augmentation de la charge du réseau. Nous décrirons les nœuds SGSN et GGSN dans la sous-section 2.1.2.

2. Ceci nécessite la présence de l'interface Iur entre deux RNCs du même réseau.

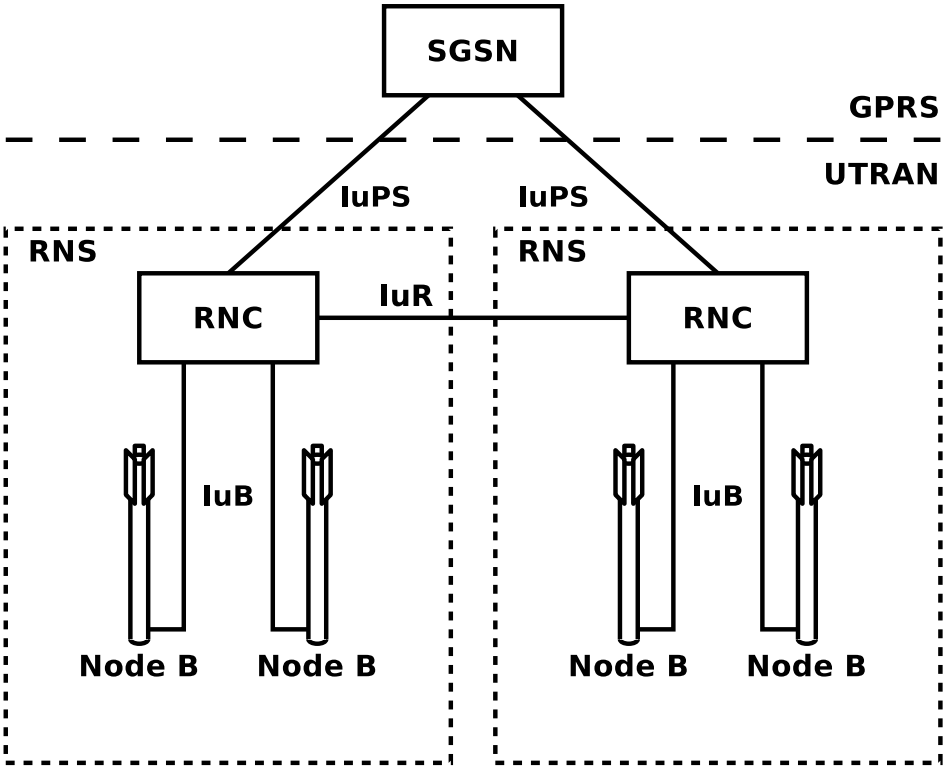


Figure 2.3 Architecture à commutation de paquets du UTRAN

Composantes du E-UTRAN

En raison des spécifications du E-UTRAN qui exigent des temps de réponse très bas sur l'interface radio, 3GPP a décidé que les fonctions traditionnellement implémentées dans le RNC le seraient dorénavant dans le LTE base station (eNodeB). En conséquence, le E-UTRAN ne contient que des eNodeBs.

La Figure 2.4 illustre le E-UTRAN. Le trafic de données transite via le Serving Gateway (S-GW) tandis que tous les messages de signalisation passent par le Mobility Management Entity (MME).[3] Cette division des tâches (qui puise ses origines dans le *Direct Tunnel Optimization*) permet aux plans de données et de contrôle d'évoluer indépendemment l'un de l'autre en fonction de l'augmentation respective de leur charge.

Les eNodeBs peuvent optionnellement être reliés à leurs voisins grâce à l'interface X2. Cette interface permet d'optimiser la procédure de relève et de transmettre les paquets que le eNodeB précédent reçoit tant que la relève n'est pas complétée.

3. Le MME sera présenté à la sous-section 2.1.2.

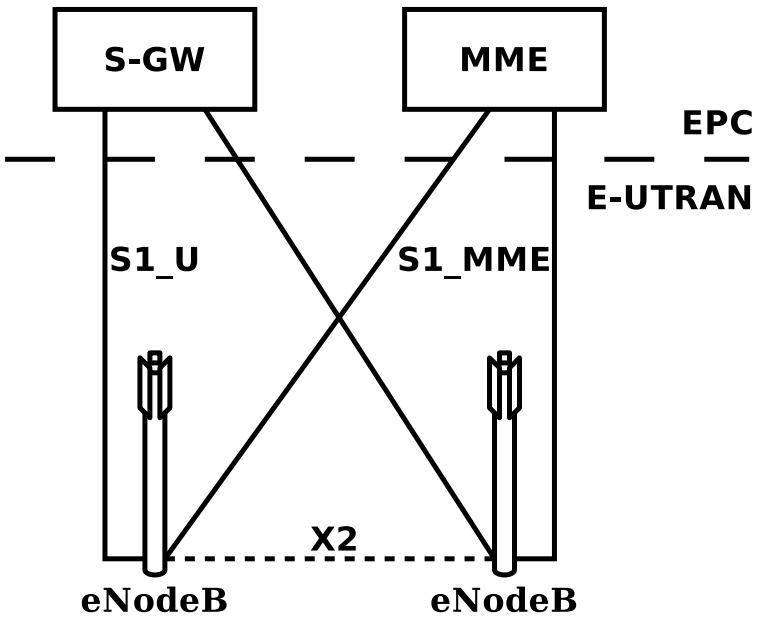


Figure 2.4 Architecture du E-UTRAN

2.1.2 Le General Packet Radio Service (GPRS)

L'architecture du réseau cœur GPRS a considérablement évolué avec l'avènement du EPS. En conséquence, l'architecture globale du EPC comprend des nœuds spécifiques aux accès 2G et 3G (se référer à la Figure 2.1 de la page 12). Nous présenterons tout d'abord les nœuds propres au EPC avant de le faire avec les nœuds des architectures 2G et 3G. Les nœuds qui ne seront pas décrits dans la présente section font partie de l'architecture PCC qui sera introduite dans la section 2.2 de la page 18.

Home Subscriber Server (HSS)

Ce nœud est une base de données centralisée qui contient les données de l'utilisateur. Le HSS puise ses origines dans l'architecture GSM puisqu'il dérive du Home Location Register (HLR) et du Authentication Center (AuC).[4] Le rôle du HLR est de conserver la dernière position connue du terminal. Celle-ci est représentée par un Tracking Area Identity (TAI). Le AuC contient les clés d'encryption qui permettent d'identifier l'usager lors de la procédure d'attachement du terminal au réseau.

4. La documentation illustre le HSS, le AuC et le SPR (vu à la sous-section 2.2.2) autant en tant qu'entités séparées que combinées. L'implémentation réelle de ces nœuds est souvent combinée en un seul équipement.

Mobility Management Entity (MME)

Le MME est une entité de contrôle cruciale au bon fonctionnement des accès 3GPP et 3GPP2. Le MME maintient la liste des terminaux en veille, supervise la création et la destruction des *bearers*, authentifie les usagers grâce au HSS/AuC. De plus, le MME doit choisir le S-GW et le P-GW (tous deux décrits ci-après) lors de la procédure d'attachement, de même que choisir le MME destination lors des relèves avec changement de MME.

Equipment Identity Register (EIR)

Le EIR est une base de données optionnelle qui permet d'identifier les terminaux déclarés perdus ou volés. Le EIR est consulté par le MME lors de la procédure d'attachement.

Cell Broadcast Center (CBC)

Le CBC est une entité logique responsable de distribuer des messages à tous les terminaux, e.g., pour qu'une agence gouvernementale distribue les alarmes liées aux événements climatiques ou sismiques. Le CBC interagit avec le MME afin que ce dernier puisse contacter les terminaux en veille.

Access Network Discovery and Selection Function (ANDSF)

Le ANDSF est une entité logique que le Mobile Network Operator (MNO) peut utiliser dans le but de contrôler comment les usagers et les terminaux priorisent la sélection des accès non-3GPP, si plusieurs technologies d'accès sont disponibles pour joindre le EPC.

Packet Data Network (PDN) Gateway (P-GW)

Le P-GW fait le lien entre EPC et un réseau IP externe, e.g., l'Internet ou un réseau corporatif. Cette entité sert d'ancrage global pour la mobilité et est aussi responsable d'assigner les adresses IP, de comptabiliser l'usage des ressources et d'émettre des Charging Data Records (CDRs). Le P-GW implémente la fonction de Policy and Charging Enforcement Function (PCEF) (voir la section 2.2). Un usager peut être connecté à plus d'un P-GW.

evolved Packet Data Gateway (ePDG)

Le ePDG fait le lien entre le EPC et les réseaux d'accès non-3GPP pour lesquels aucune relation de confiance n'a été établie. Ce nœud fait entre autres le relai de la signalisation permettant d'authentifier l'usager et le terminal. De plus, le ePDG termine le lien encrypté qui a été établi avec le terminal.

Serving Gateway (S-GW)

Le S-GW est l'entité logique à la frontière du RAN et du EPC qui achemine le trafic des utilisateurs. Il sert d'ancrage à la mobilité inter-eNodeB de même que pour les relèves inter-3GPP (i.e., entre les GERAN, UTRAN et E-UTRAN). Une autre fonction importante du S-GW est de stocker temporairement les données destinées aux terminaux en mode de veille pendant que le MME les invitent à se reconnecter au réseau. Un terminal ne peut être connecté qu'à un seul S-GW à la fois.

Dans la variante PMIPv6, le S-GW doit remplir des fonctions supplémentaires :

- rapporter les événements radio au PCRF (décrit à la section 2.2) ;
- effectuer le *bearer binding*[5] ;
- jouer le rôle de Mobility Access Gateway (MAG) selon la spécification PMIPv6 ;
- gérer les messages *Router Sollicitation*, *Router Advertisement*, *Neighbor Sollicitation*, *Neighbor Advertisement* ;
- jouer le rôle de *relay agent* pour le Dynamic Host Configuration Protocol (DHCP), tant pour DHCPv4 que pour DHCPv6 ;
- allouer les clés en aval de l'encapsulation Generic Routing Encapsulation (GRE) utilisée pour distinguer les bearers des usagers.

Serving GPRS Support Node (SGSN)

Le SGSN est une entité logique des RANs 2G et 3G. Dans un réseau cœur GSM ou UMTS, il est approximativement l'équivalent de la combinaison du MME et du S-GW.[6] Le SGSN agit à titre d'ancrage à la mobilité pour l'ensemble du GERAN et du UTRAN.

Au sein du EPS, le SGSN interagit avec le MME pour le contrôle des terminaux lorsque ceux-ci sont connectés au GERAN ou au UTRAN (e.g., en absence de couverture LTE).

Pour les usagers ayant un terminal multimode avec abonnement LTE, le trafic de ces terminaux passe par le S-GW.[7] Dans le cas contraire, le trafic des usagers transite par le GGSN (décrit ci-après). Le SGSN peut déterminer à quel nœud (entre le GGSN ou le S-GW) envoyer le trafic d'un usager en fonction de l'identificateur du réseau à joindre.[8]

5. Le *bearer binding* est l'association des Service Data Flows (SDFs) aux tunnels GTP que l'on retrouve entre les eNodeBs et le S-GW.

6. Les fonctions de contrôle et de transport de données ont été séparées dans le EPC dans le but de permettre aux plans de données et de contrôle d'évoluer indépendamment selon leurs besoins respectifs.

7. Lorsque le *Direct Tunnel Optimization* est utilisé, le trafic de l'usager passe outre le SGSN pour se rendre directement au S-GW.

8. Par exemple, pour se connecter à l'Internet, un usager ayant droit à l'accès LTE utiliserait un Access Point Name (APN) différent d'un usager 2G/3G.

Gateway GPRS Support Node (GGSN)

Le GGSN est une entité logique des architectures 2G/3G dont le rôle est d'interconnecter le réseau GPRS à un réseau IP externe et agir à titre d'ancrage à la mobilité. Le GGSN alloue les adresses IP aux terminaux et comptabilise l'utilisation des ressources. De plus, il est responsable de gérer la QoS et met en application les politiques du réseau.

2.2 Architecture PCC du EPS

L'architecture PCC implémente les fonctions de gestion des politiques du réseau ainsi que la collecte d'informations sur l'utilisation des ressources. Dans cette section, nous décrirons chacun des nœuds logiques de cette architecture (voir la Figure 2.5). Les éléments en pointillé sont spécifiques à la variante PMIPv6 du EPC. Il existe trois scénarios :

- sans itinérance (*non-roaming*) ;
- itinérance avec trafic encapsulé vers le réseau de l'usager (*home-routed*) ;
- itinérance avec trafic localisé dans le réseau visité (*local breakout*).

2.2.1 Application Function (AF)

Le AF représente un serveur d'application qui interagit avec l'architecture PCC afin d'obtenir la QoS nécessaire à son bon fonctionnement. Lors d'un scénario d'itinérance, les AFs peuvent être situés dans le réseau de l'usager ou dans le réseau visité.

2.2.2 Subscription Profile Repository (SPR)

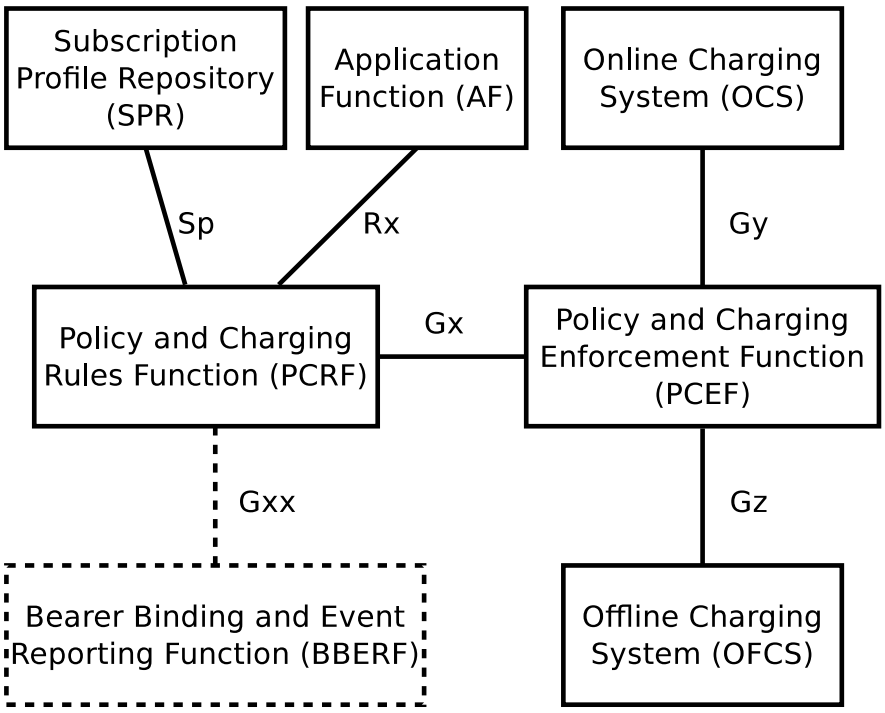
Le SPR est une base de données qui contient les profils d'application des usagers, soit la liste des services auxquels ils ont souscrit ainsi que certains paramètres de QoS.

2.2.3 Policy and Charging Rules Function (PCRF)

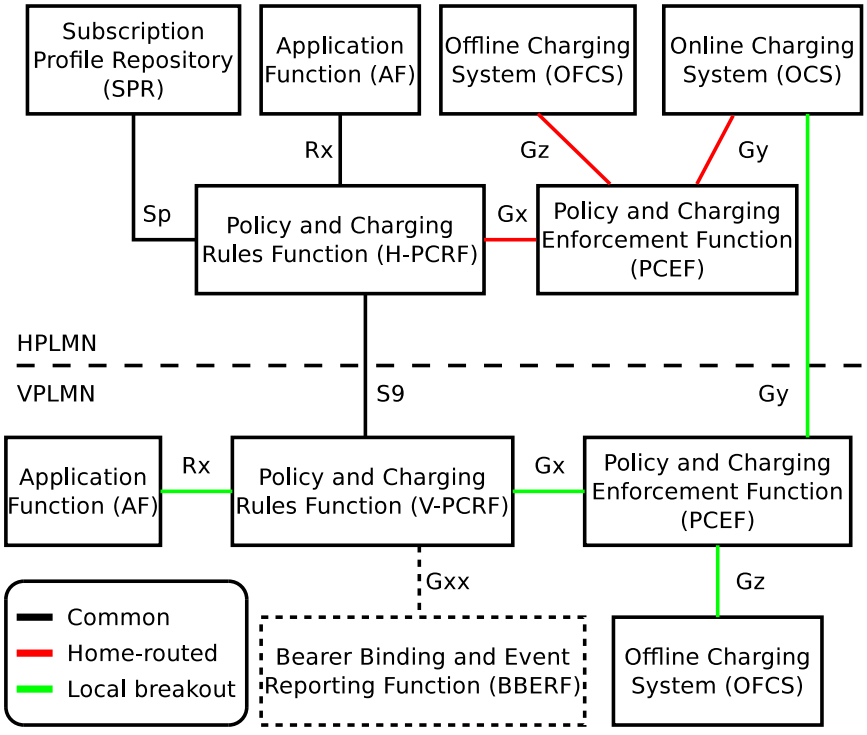
Le PCRF est une entité logique qui décide des politiques à appliquer en fonction de l'identité de l'usager qui émet une requête de QoS, son profil d'application, le type de service invoqué de même que les conditions actuelles du réseau.

2.2.4 Policy and Charging Enforcement Function (PCEF)

Le PCEF est une entité logique qui applique les politiques statiques contenues à même sa configuration ou celles énoncées par le PCRF. Le PCEF identifie les flots de données, contrôle leur débit, marque les paquets (QoS), effectue le *bearer binding* (variante GTP seulement). Enfin, le PCEF émet des billets de consommation servant ultimement à la facturation.



(a) Architecture sans itinérance.



(b) Cas d'itinérance *Home-routed* et *local breakout*.

Figure 2.5 Architecture PCC logique (avec ou sans itinérance)

2.2.5 Bearer Binding and Event Reporting Function (BBERF)

Le BBERF est une entité logique qui n'existe que dans la variante PMIPv6 du EPC et qui est implémentée dans tous les Access Edge Gateways (AEGs). [9] En effet, le protocole PMIPv6 ne gère que la mobilité tandis que GTP gère en plus la QoS et les procédures des *bearers*. Le BBERF effectue le *bearer binding* et rapporte les événements radio au PCRF.

2.2.6 Offline Charging System (OFCS)

Le OFCS est une entité logique qui traite les CDRs provenant des éléments du réseau seulement après que les ressources furent consommées. Le rôle principal du OFCS est d'interagir avec le mécanisme de facturation du MNO.

2.2.7 Online Charging System (OCS)

Le OCS est une entité logique qui effectue un traitement semblable à celui du OFCS mais pour la consommation de ressources prépayées. Des traitements supplémentaires sont requis pour authoriser les ressources réseau avant leur consommation à proprement parler.

2.3 Mécanismes de QoS pour les technologies d'accès

Cette section présente un survol des technologies d'accès qui peuvent se connecter au EPC. Nous porterons une attention particulière au paradigme de QoS utilisé (réservation ou différentiation) sur le lien radio, au mécanisme que les terminaux utilisent pour accéder au canal (avec ou sans compétition) de même qu'à l'initialisation de la QoS sur le lien radio (par le terminal ou par le réseau).

2.3.1 Les accès 3GPP

Pour tous les accès 3GPP, l'utilisation de la QoS dans le réseau GPRS est basée sur une combinaison de réservation de ressources et de différentiation de services. En effet, des *bearers* (on utilise le terme « *PDP context* » pour GSM et UMTS) ayant leurs propres paramètres de QoS relient le terminal au RAN.

Suite à la procédure d'attache, le terminal (le réseau dans le cas du E-UTRAN) crée un *default bearer* [10] dont le débit n'est pas garanti et qui offre une connectivité de base à celui-ci. Tous les autres *bearers* sont dits dédiés. [11]

9. Pour les accès 3GPP, le BBERF est implémenté dans le S-GW.
10. On utilise le terme « *Primary PDP context* » pour GSM et UMTS.
11. GSM et UMTS utilisent le terme « *Secondary PDP context*. »

Dans le E-UTRAN, les *bearers* dédiés se divisent en deux catégories :

Guaranteed Bit Rate (GBR) : ceux-ci possèdent leurs propres ressources réservées et leur débit est garanti ;

non-GBR : ceux pour lesquels une différentiation de services est appliquée afin de distribuer la bande passante disponible.

La somme des débits des *bearers* non-GBR est soumise à des limites (distinctes en amont et en aval [12]) pour le terminal [13] ainsi que pour le réseau auquel celui-ci est connecté. [14] En conséquence, ces quatre paramètres suivants constituent les Aggregate Maximum Bit Rates (AMBRs) : UE-DL-AMBR, UE-UL-AMBR, APN-DL-AMBR, APN-UL-AMBR.

Pour les accès GSM et UMTS jusqu'à la version 6 inclusivement de l'architecture 3GPP, la création d'un *secondary PDP context* était exclusivement la responsabilité du terminal. La version 7 a été amendée de telle sorte qu'une procédure de création/modification de bearer soit ajoutée du côté réseau. Enfin, le EPS (versions 8 et suivantes) contrôle entièrement les bearers (le terminal peut encore demander des ressources mais c'est toujours le réseau qui prend une décision finale) [45].

Enfin, tous les échanges de paquets entre les terminaux et la station de base sont régis par le RAN ; il n'y a donc jamais (à l'exception de la procédure d'attache) de compétition pour accéder au canal radio.

2.3.2 Les accès 3GPP2

Les accès de la famille 3GPP2 sont très similaires aux technologies 3GPP 2G et 3G correspondantes, en termes de performance et de fonctionnement.

Pour les besoins de notre recherche, les aspects suivants sont à quelques détails techniques près identiques pour les architectures proposées par 3GPP et 3GPP2 :

- l'architecture générale du réseau incluant la distribution des fonctions parmi les nœuds logiques de celle-ci ;
- les procédures de gestion des bearers radio ;
- le nombre ainsi que les caractéristiques des classes de QoS ;
- le fait que le terminal soit responsable de créer les bearers radio ;
- le fait que l'accès au canal radio soit contrôlé par le RAN.

De plus, 3GPP et 3GPP2 ont défini les mécanismes servant à interfacer un RAN 3GPP2 avec un réseau cœur 3GPP [7, 8]. Finalement, il n'y a plus de technologie 3GPP2 en compétition directe avec LTE. Effectivement, 3GPP2 a retiré en 2008 de sa liste de spécifications

12. En anglais, les termes « Downlink (DL) » et « Uplink (UL) » désignent respectivement l'amont et l'aval.
13. Le terme « User Equipment (UE) » désigne le terminal en terminologie 3GPP.
14. Le terme « Access Point Name (APN) » désigne le réseau du MNO en terminologie 3GPP.

la technologie Ultra Mobile Broadband (UMB) en faveur de LTE pour l'évolution du réseau d'accès 3GPP2 à commutation de paquets (l'équivalent de GPRS), connu sous le nom de High Rate Packet Data (HRPD).

2.3.3 IEEE 802.11 avec support de la QoS

Les points d'accès 802.11 [28] sont omniprésents dans les résidences et les bureaux. Ils peuvent décharger les réseaux cellulaires lors des situations de faible mobilité des terminaux. [15] De nos jours, peu de points d'accès supportent la QoS telle que définie dans le standard.

Les stations 802.11 peuvent constituer des réseaux locaux en modes *ad hoc* ou *infrastructure*. Le mode *ad hoc* implique que les stations échangent des données sans la coordination d'une entité centrale et ne présente aucun intérêt pour cette recherche. En contrepartie, le mode *infrastructure* implique la présence d'un Access Point (AP) qui gère le canal radio et par lequel transitent généralement toutes les trames. [16] Dans cette recherche, nous ne considérerons que le mode *infrastructure*.

Il existe quatre mécanismes d'accès au canal. Le mécanisme fondamental, supporté par toutes les stations et les APs, est connu sous le nom de Discrete Coordination Function (DCF). Il s'agit d'une implémentation du mécanisme d'accès réparti Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). Lorsqu'une station veut émettre, celle-ci doit attendre un intervalle de temps nommé DCF InterFrame Space (DIFS) (lorsqu'elle perçoit que le canal devient libre) avant de décrémenter un compteur additionnel (servant à limiter le nombre potentiel de collisions). Si le canal est toujours libre suite à cette attente, la station est autorisée à émettre un court message Request To Send (RTS) auquel le AP répond par un message Clear To Send (CTS). Dès que les autres stations captent le RTS ou le CTS, celles-ci doivent garder le silence jusqu'à la fin de la transmission. Le mécanisme DCF est utilisé lors des Contention Periods (CPs) lors desquelles des collisions peuvent survenir.

Un second mécanisme permet à un AP de prendre le contrôle du canal radio de façon exclusive et ainsi éviter les collisions : on parle alors des Contention-Free Periods (CFPs). Le mécanisme est connu sous le nom de Point Coordination Function (PCF). Le AP prend le contrôle du canal (typiquement 10 fois par seconde) en questionnant à tour de rôle toutes les stations enregistrées au service PCF afin de leur permettre d'envoyer des données de même que recevoir les trames accumulées au niveau du AP. [17] Lorsque le AP a questionné toutes les stations, le mécanisme DCF est utilisé jusqu'à la prochaine CFP.

15. Des études montrent qu'environ 70% des appels 3G sont initiés de l'intérieur d'un édifice [19].

16. Notez que la norme actuelle [28] définit le mécanisme Direct Link Setup (DLS) pour l'échange direct de trames entre deux stations voisines.

17. Le AP a un accès privilégié au canal car ce dernier ne doit attendre qu'un intervalle de temps PCF InterFrame Space (PIFS), qui est plus court qu'un DIFS, pour accéder au canal radio.

Une caractéristique importante des canaux radio en 802.11 est que ces derniers sont utilisés autant pour les transferts en amont qu'en aval. De plus, les modes DCF et PCF ne peuvent supporter la QoS car ceux-ci ne possèdent aucun moyen de contrôler le temps qu'une station conserve le canal [40]. En effet, si une station éloignée doit recourir à une modulation à faible débit et transmet une longue trame, celle-ci peut occuper le canal pendant un intervalle de temps inacceptable pour satisfaire les besoins en QoS des autres stations.

Les deux derniers mécanismes d'accès au canal, EDCA et HCCA, sont respectivement des variantes de DCF et PCF offrant un bien meilleur support de la QoS. L'unité fondamentale d'allocation de ressources du canal est une tranche de temps connue sous le nom de Transmission Opportunity (TXOP). Lorsqu'une station obtient l'accès exclusif au canal, celle-ci peut transmettre plusieurs trames mais sans dépasser le TXOP. Ainsi, quelle que soit la modulation radio utilisée pour le transfert des trames, le canal radio ne sera jamais occupé trop longtemps.

Le Enhanced Distributed Channel Access (EDCA) est un mécanisme d'accès distribué qui améliore DCF par l'utilisation des TXOPs de même que par la définition de quatre classes de trafic nommées Access Categories (ACs). Ces dernières permettent la différentiation de services et sont directement associées aux classes de services définies dans la norme IEEE 802.1D [25]. Les ACs sont définies par un niveau de priorité d'accès au canal, les limites inférieure et supérieure d'attente supplémentaire avant d'accéder au canal de même que par le TXOP. EDCA est utilisé pendant les CPs en remplacement de DCF. Le AP peut spécifier au début de chaque cycle d'échanges (*superframe*) dans un message diffusé à tous (*beacon*) les durées des TXOPs pour chaque AC.

Enfin, le HCF Controlled Channel Access (HCCA) est un mécanisme d'accès centralisé très avancé qui améliore PCF en allouant des TXOPs aux stations de façon dynamique, selon les ACs des flots à supporter de même que de la charge du AP. En effet, le AP peut débuter à tout moment une CFP si les besoins en QoS des flots définis le justifient. HCCA est basé sur des requêtes de réservations de ressources initiées par les stations. La QoS en UL et en DL est distincte. Suite à une requête de QoS, le AP est libre de l'accepter ou de la refuser, selon une décision de politique de réseau.

2.3.4 IEEE 802.16-2009

Le standard IEEE 802.16 actuel (publié en 2009) permet à une station mobile d'effectuer des relèves, et de ce fait a transformé l'interface radio 802.16 précédente (de 2004) afin de lui procurer les caractéristiques d'un réseau cellulaire. En effet, la version de 2004 de la norme avait défini un accès fixe sans-fil et à large bande. L'amendement 802.16e publié en 2005 a permis d'incorporer le concept de mobilité.

La station de base alloue des tranches de temps du canal radio pour le trafic aussi bien en UL qu'en DL : ceci permet d'assurer la QoS et élimine les risques de collision de trames. Un total de cinq classes de services sont définies pour la création des SDFs. De plus, la réservation de ressources au niveau de l'interface radio peut être effectuée par le client mobile ou le réseau. Toutefois, tout comme 3GPP pour le EPS, le consortium WiMAX démontre une forte préférence pour la QoS activée par le réseau.

Il existe deux modèles de QoS définis pour l'architecture WiMAX :

- Le *modèle Authentication, Authorization and Accounting (AAA)* selon lequel le profil de QoS de l'usager est stocké dans le serveur AAA ; ce dernier incorpore un mécanisme dynamique de prise de décisions de politiques ;
- le *modèle PCC* selon lequel le réseau d'accès WiMAX est branché à un réseau cœur 3GPP ; l'architecture PCC du EPC effectue donc ce travail.

2.4 Tendances de l'industrie

Le monde des communications mobiles a considérablement évolué depuis ses débuts et a atteint un haut degré de maturité. En conséquence, les opérateurs sont maintenant entrés dans une phase de réduction de coûts au détriment de l'innovation. De plus, le processus de maturation encourage l'émergence de nouveaux opérateurs qui visent des clientèles de plus en plus spécialisées (e.g., étudiants, gens d'affaires, retraités). Enfin, la forte intégration verticale des opérateurs est caractéristique de leur immaturité ; avec l'augmentation de la complexité, aucune organisation ne peut offrir à elle seule une solution complète [50].

L'évolution des technologies d'accès sans-fils se poursuit à un rythme de plus en plus rapide et gagne en complexité. Parallèlement, les lois d'un pays peuvent imposer aux opérateurs d'offrir une couverture dans des zones qui sont peu attrayantes d'un point de vue commercial. En conséquence, les opérateurs doivent mieux contrôler leurs coûts afin de demeurer compétitifs et donc se concentrer sur les activités avec une grande valeur ajoutée [20].

Afin de réduire leurs coûts d'opération, un nombre grandissant d'opérateurs sous-traitent les activités les moins payantes ou sont encouragés à partager certaines de leurs infrastructures physiques. Le revenu moyen par usager est la principale motivation de ces actions puisque ce dernier est demeuré à peu prêt constant alors que les volumes de voix et de données augmentent constamment [53].

Le partage d'infrastructures est intéressant autant pour les opérateurs déjà établis que pour les nouveaux joueurs dans ce marché. En effet, les opérateurs établis ont l'occasion de générer de nouveaux revenus tandis que les MVNOs émergents peuvent très rapidement accéder à un immense territoire sans devoir subir le fardeau de déployer seuls un RAN.

Par ailleurs, des opérateurs ayant des clientèles fort différentes bénéficient d'un avantage additionnel en partageant leurs infrastructures : le dimensionnement des réseaux dépend beaucoup de la demande maximale et les pics de consommation peuvent survenir à des moments différents de la journée. Cela a pour conséquence d'augmenter la charge moyenne du réseau sans augmenter significativement la charge maximale à supporter [11, 50].

Alors que se produisent les transformations citées précédemment, les opérateurs disposant de réseaux d'accès fixe et mobile ont réalisé qu'il serait possible de réduire davantage leurs coûts d'exploitation en connectant les deux réseaux d'accès au même réseau cœur. De plus, l'apparition de terminaux multimode permet à ces opérateurs d'offrir à leurs clients soit la meilleure connexion possible ou celle dont le coût par bit est le plus avantageux [53]. Enfin, puisqu'il y de nombreux avantages à la FMC, les opérateurs exclusivement fixes ou mobiles sont poussés à bâtir des alliances afin de demeurer compétitifs.

Compte tenu des tendances de l'industrie, les auteurs de [11] prévoient que les opérateurs de réseaux se consolideront et s'efforceront de réduire leurs coûts alors que les fournisseurs de services se fragmenteront et tenteront d'innover à tout prix. Ainsi, quelques grands fournisseurs d'infrastructures physiques devraient résulter de ces transformations. Chacun d'eux sera à l'origine d'un ou plusieurs fournisseurs d'accès au réseau. Il devrait y avoir des centaines de fourniseurs de services qui supportent des milliers de services (Figure 2.6).

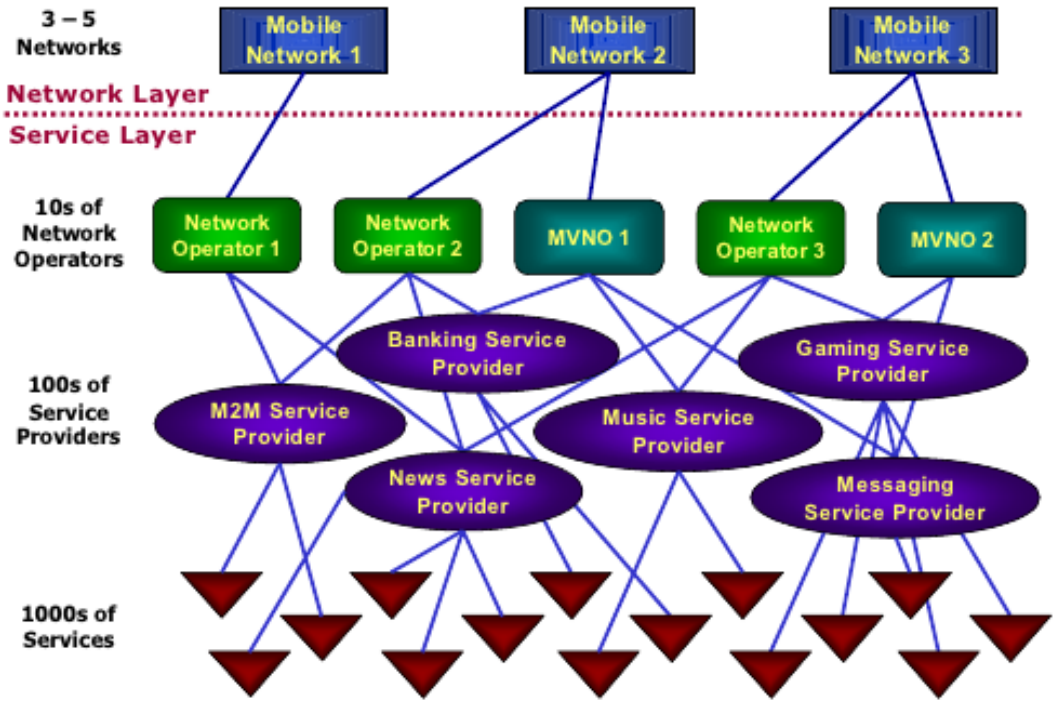


Figure 2.6 Aperçu de la structure de l'industrie des communications mobiles. *Source : [50]*

2.5 Revue de littérature sur le partage d'infrastructures

Dans cette section nous présenterons un ensemble de publications qui portent sur le sujet du partage d'infrastructures physiques entre plusieurs opérateurs. Toutefois, nous ne présenterons aucune publication qui traite exclusivement de partage passif des infrastructures [18] ou des scénarios de partage de stations de base et/ou de RNC pour lesquels chaque MNO a ses bandes de fréquences dédiées. De même, les articles qui traitent d'optimisation du partage des ressources radio de même que ceux qui proposent des modèles économiques ayant pour but d'optimiser ou modéliser les coûts de déploiement des infrastructures sont considérés hors sujet. Enfin, la majorité des articles suivants proposent la virtualisation comme solution au partage des infrastructures entre plusieurs MNOs.

La virtualisation a pour but premier de réduire les coûts d'exploitation en permettant le partages des infrastructures entre plusieurs opérateurs tout en garantissant l'isolation et l'indépendance des opérateurs entre les réseaux virtuels. Puisque le potentiel d'économies est élevé, le partage d'infrastructures a d'abord été considéré pour le RAN [24].

2.5.1 Partage d'infrastructures dans les réseaux 3GPP actuels

Le partage passif du RAN n'a pas besoin d'être standardisé car chaque MNO est maître de ses bandes de fréquences (chaque canal n'annonce qu'un seul MNO) ; seuls les éléments du RAN supportent une forme de virtualisation afin de séparer les opérateurs.

Le partage actif d'infrastructures implique que les bandes de fréquences des opérateurs soient mises en commun (le terme utilisé est *frequency pooling*). Les stations de base annoncent la liste de tous les MNOs disponibles et les UEs qui supportent le partage peuvent choisir leur MNO tandis que les autres se font assigner un MNO par le RAN.

Actuellement, seules deux approches au partage actif d'infrastructures existent [3] :

Multi-Operator Core Network (MOCN) où seuls les éléments du RAN sont partagés (i.e., RNCs et Node Bs). Ainsi, chaque MNO possède son propre réseau cœur.

Gateway Core Network (GWCN) où le RAN et les quelques éléments du réseau cœur (e.g., SGSNs, MMEs) sont partagés.

2.5.2 Partage d'infrastructures pour le projet 4WARD

Le projet européen 4WARD élabore ce à quoi devrait ressembler un futur Internet en termes de QoS, virtualisation des ressources, partage d'infrastructures, gestion du réseau, etc. Le but visé étant que le réseau devrait être en mesure de mieux s'autogérer et automatiser

18. Le partage d'infrastructures est dit *passif* si aucune collaboration entre les MNOs n'est requise, e.g., le partage d'abris, de mâts et d'antennes. Il est dit *actif* dans le cas contraire.

de nombreuses activités de gestion. De plus, la gestion du réseau devrait être fortement décentralisée, contrairement à ce qui se fait de nos jours.

La virtualisation constitue l'un des sous-projets importants de 4WARD. Le but est de faire progresser la virtualisation des couches 2 et 7 (liens et applications, tel que fait actuellement) vers les couches 3 et 4 (réseau et transport). Ainsi, il deviendra possible pour plusieurs architectures différentes de coexister, grâce à la séparation créée entre les services et les infrastructures physiques. De plus, cette coexistence peut réduire considérablement le temps requis pour passer à une nouvelle architecture car les participants n'ont plus nécessairement besoin d'établir un large consensus concernant le déploiement d'une technologie donnée. En conclusion, il devient possible de faire coexister les réseaux actuels tout en permettant le déploiement rapide de nouveaux protocoles, services et architectures.

L'ensemble des problèmes de virtualisation (Figure 2.7) qui seront abordés dans le projet 4WARD sont divisés en trois principaux domaines [9] :

Virtualisation des ressources du réseau : Les ressources incluent les liens et nœuds virtuels, espaces de stockage ou CPUs, et appartiennent à un fournisseur d'infrastructures physiques. Des interfaces normalisées permettront de gérer les ressources virtuelles.

Création de réseaux virtuels : Développement d'un cadre pour découvrir les ressources physiques et virtuelles, basé sur une approche évolutive afin de fournir, contrôler et agréger des ressources, dans le but de constituer des réseaux virtuels complets.

Gestion de la virtualisation : Des mécanismes de gestion à la demande permettront de déployer, contrôler et réallouer dynamiquement des ressources, tout au cours de la vie du réseau virtuel.

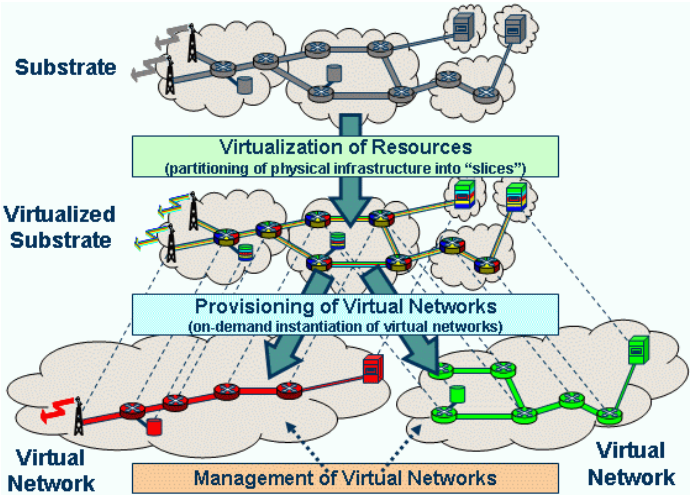


Figure 2.7 Concepts de virtualisation du projet 4WARD. *Source : Projet 4WARD*

2.5.3 Partage d'infrastructures avec GENI

Le projet américain Global Environment for Network Innovations (GENI) a pour but de créer un cadre de travail permettant aux scientifiques et ingénieurs de conduire des expériences en réseautique sur des réseaux existants qui sont partagés entre les parternaires du projet [21].

La virtualisation des ressources du réseau rend possible un découpage de ces dernières en tranches (*slices*) de façon à isoler les expériences les unes des autres. Les tranches de réseau sont gérées à l'aide d'OpenFlow.

2.5.4 Isolation de flots avec OpenFlow

Le concept de réseau programmable est à l'origine d'OpenFlow. En effet, OpenFlow consiste en un module logiciel que l'on ajoute aux commutateurs et routeurs dans le but de gérer des filtres de paquets qui nécessitent un traitement particulier.

L'objectif principal d'OpenFlow est de faciliter l'expérimentation de nouveaux protocoles sur des équipements de réseaux existants (routeurs et commutateurs), et ce sans affecter le fonctionnement normal de ces équipements. Grâce à une interface normalisée, il devient possible de gérer des filtres de flots de façon uniforme sur des équipements hétérogènes [41].

L'architecture d'OpenFlow est simple : un contrôleur centralisé interagit avec un ou plusieurs commutateurs OpenFlow. Le contrôleur gère l'ensemble des filtres de flots qui sont configurés dans les commutateurs. Les filtres contiennent 10 champs qui permettent de capturer des paquets aux niveaux 2 à 4 inclusivement.

Un minimum de trois actions doivent ètre supportées par les commutateurs OpenFlow :

1. retransmission du paquet vers un port donné ;
2. encapsulation et envoi du paquet au contrôleur ;
3. élimination du paquet.

FlowVisor

OpenFlow peut supporter plus d'un contrôleur mais ces derniers seront surtout utilisés pour offrir de la redondance ou pour améliorer les performances. S'il faut permettre à plusieurs expériences d'exécuter en parallèle, FlowVisor [49] constitue une solution simple et élégante. En effet, FlowVisor peut créer plusieurs tranches dans le réseau et déléguer chacune de ces tranches à un contrôleur différent.

FlowVisor agit comme un serveur proxy transparent qui intercepte tous les messages OpenFlow entre les commutateurs et les contrôleurs. De plus, il peut aussi intercepter et modifer les messages d'un autre FlowVisor. Ainsi, FlowVisor agit comme un gestionnaire de ressources et un serveur de politiques.

2.5.5 Méthodes traditionnelles d'isolation de flots

L'isolation de flots n'est pas un concept récent en réseautique. Toutefois, les méthodes traditionnelles ne se concentrent qu'aux niveaux 1 à 3. En conséquence, les seules ressources partageables sont des liens physiques, des tranches de temps, des longueurs d'ondes, etc.

Par ailleurs, la configuration et la gestion des mécanismes traditionnels doivent être effectuées par une seule entité de contrôle, e.g., le fournisseur d'infrastructures physiques. Ainsi, les clients ne disposent généralement pas de mécanismes permettant de réserver ou modifier à la demande des ressources du réseau.

Un premier mécanisme d'isolation de trafic est le circuit virtuel. Un circuit virtuel est un tunnel logique qui transporte des données (avec ou sans QoS) à travers un réseau à commutation de paquets et de façon transparente pour les données transportées. Des exemples de telles technologies sont Frame Relay (FR) et Asynchronous Transfer Mode (ATM).

Les réseaux Ethernet locaux

Les réseaux locaux Ethernet supportent un partitionnement grossier du trafic grâce aux Virtual Local Area Networks (VLANs), définis dans le norme 802.1Q [26]. L'étiquette VLAN est destinée à isoler le trafic provenant des commutateurs aux frontières de la dorsale. [19] Les VLANs peuvent par exemple isoler le trafic de deux départements au sein d'un entreprise.

Lorsqu'une trame arrive à la frontière du réseau (qui sépare la dorsale du "client"), le commutateur ajoute un champ de 32 bits à l'en-tête Ethernet pour spécifier le VLAN ainsi que le niveau de priorité de la trame. Si l'adresse Medium Access Control (MAC) destination est inconnue, le commutateur ne doit retransmettre en *broadcast* la trame que sur les ports qui font partie du même VLAN. [20]

Une amélioration au concept de VLAN connue sous le nom de Q-in-Q est définie dans la norme 802.1ad [27]. Un commutateur peut insérer un en-tête VLAN, défini pour la dorsale, devant un en-tête VLAN défini par le client. Ainsi, une meilleure séparation des clients et de la dorsale est atteinte.

Notez que les normes 802.1Q et 802.1ad ne sont pas applicables aux grands réseaux car les adresses MAC des trames proviennent des réseaux locaux connectés à la dorsale. En conséquence, le mécanisme d'apprentissage des adresses MAC des commutateurs ne serait pas évolutif puisqu'il lui faudrait apprendre toutes les adresses MAC des clients.

19. Par exemple, les trames émises en *broadcast* en provenance d'un VLAN ne doivent en aucun cas être retransmises aux autres VLANs.

20. Pour fin d'optimisation, il est possible que l'arbre de recouvrement des commutateurs, normalement déterminé par le Rapid Spanning Tree Protocol (RSTP), soit évalué pour chaque VLAN individuellement grâce à l'ajout du Multiple Spanning Tree Protocol (MSTP) dans la norme 802.1Q de 2005.

Les réseaux Ethernet métropolitains

Les réseaux Ethernet sont de plus en plus présents dans les réseaux métropolitains en raison de leurs faibles coûts et de leur omniprésence dans la majorité des réseaux locaux.

Suite au problème d'évolutivité décrit pour les normes 802.1Q et 802.1ad, la variante 802.1ah [21] [29] d'Ethernet fut standardisée. Sa principale caractéristique est d'encapsuler la trame en provenance d'un client à l'intérieur d'une autre trame Ethernet. Ainsi, les commutateurs de la dorsale n'ont plus accès aux adresses MAC des trames des clients, ce qui permet le déploiement à grande échelle de cette technologie.

Pour les réseaux dorsaux qui doivent être déterministes et offrir une grande résistance aux pannes, la norme 802.1Qay [22] [30] utilise le même format de trame que le standard 802.1ah mais a éliminé les mécanismes de *flooding* (lorsque l'adresse destination d'une trame n'est pas connue), d'apprentissage dynamique des adresses MAC ainsi que le Spanning Tree Protocol (STP). Il en résulte une dorsale pour laquelle les commutateurs sont configurés statiquement mais qui supportent dorénavant l'ingénierie de trafic [23] de même qu'être interopérable avec la technologie Multi-Protocol Label Switching (MPLS). Enfin, des mécanismes de récupération permettent la détection et la correction d'une panne en moins de 50 ms.

MPLS et ses variantes

Nous avons mentionné les technologies FR et ATM pour la création de circuits virtuels. Ces derniers sont utilisés pour séparer des flots, garantir une QoS et s'assurer de rencontrer les objectifs de performance du réseau au moyen de l'ingénierie de trafic.

Multi-Protocol Label Switching (MPLS) [48] est un mécanisme permettant à un réseau IP de créer des circuits virtuels [24] basés sur la commutation d'étiquettes, d'où le nom de Label-Switched Path (LSP). Les commutateurs MPLS (qui sont aussi des routeurs IP) portent le nom de Label-Switching Router (LSR). Les Label Edge Routers (LERs) sont des LSRs situés aux frontières du domaine MPLS qui ont pour responsabilité d'associer au LSP correspondant les paquets arrivant de l'extérieur du domaine MPLS.

21. Aussi connue sous le nom de *Provider Backbone Bridge (PBB)*.
22. Aussi connue sous le nom de *Provider Backbone Bridge-Traffic Engineering (PBB-TE)*.
23. L'ingénierie de trafic est essentielle à la gestion des réseaux modernes car elle permet l'atteinte d'objectifs visant à rentabiliser les équipements très coûteux tout en offrant une QoS aux utilisateurs.
24. Puisque MPLS est indépendant de la technologie de transport (aux couches 1 et 2) et qu'il peut transporter des paquets IP, on dit souvent que c'est un mécanisme de la couche 2.5.

Pour les réseaux destinés aux grands transporteurs en télécommunications, une variante connue sous le nom de MPLS Transport Profile (MPLS-TP) est basée sur le plan de données de MPLS mais retire le support pour deux mécanismes jugés superflus :

1. le Penultimate Hop Popping (PHP) qui permet à l'avant-dernier LSR de retirer l'en-tête MPLS avant d'acheminer le paquet vers le LER de sortie ;
2. le Equal Cost MultiPath (ECMP) qui permet de créer un LSP virtuel combinant plusieurs LSPs parallèles afin d'en augmenter le débit.

Enfin, contrairement à MPLS, MPLS-TP réserve une étiquette pour créer un canal Operations, Administration and Maintenance (OAM) à même chaque LSP. Ce canal (nommé *Generic Associated Channel*) permet entre autres aux LSRs de détecter et prendre action afin de corriger une panne en moins de 50 ms.

La gestion dynamique des LSPs est faite à l'aide du Label Distribution Protocol (LDP) lorsque la QoS n'est pas considérée. Dans le cas contraire, deux protocoles de réservation de ressources et de distribution d'étiquettes sont supportés :

1. Constraint-Routing Label Distribution Protocol (CR-LDP) ;
2. Resource reSerVation Protocol for Traffic Engineering (RSVP-TE). [25]

Les concepts de commutation d'étiquettes de MPLS peuvent être généralisés afin de s'appliquer à des paramètres physiques ou logiques des liens. En effet, les différentes longueurs d'ondes, le multiplexage temporel (e.g., SONET/SDH) de même que les considérations spatiales (e.g., les ports d'entrée et de sortie d'un commutateur) deviennent autant de types de commutations pouvant influencer le trajet emprunté par une trame. [26]

L'objet d'intérêt de Generalized MPLS (GMPLS) [34] est le plan de contrôle pour tous les paramètres mentionnés ci-haut puisque chacun d'eux peut utiliser un plan de données physiquement différent. GMPLS permettra de contrôler la signalisation et le routage sur ce plan de contrôle unifié. En effet, l'architecture originale de MPLS fut étendue afin d'inclure les LSRs dont les plans de données ne reconnaissent pas les frontières des paquets ou cellules, et ainsi ne peuvent effectuer la commutation sur des informations contenues dans les en-têtes des paquets ou cellules.

25. Suite à une décision de l'IETF [10], CR-LDP ne sera plus étendu. Seul RSVP-TE sera supporté.

26. Par exemple, une fibre optique peut transporter plusieurs canaux en parallèles grâce aux différentes longueurs d'ondes supportées. Si l'on utilise l'information de la longueur d'onde comme étiquette il devient alors possible pour un commutateur optique de choisir le chemin qu'une trame doit prendre, et ce en fonction de la longueur d'onde de la trame sur la fibre entrante.

En conséquence, de nouvelles classes d'interfaces de LSRs sont introduites par GMPLS :

Packet Switch Capable (PSC) **:** interfaces qui reconnaissent les frontières d'un paquet et peuvent interpréter les informations de son en-tête ;

Layer-2 Switch Capable (L2SC) **:** interfaces qui reconnaissent les frontières d'une cellule ou d'une trame et peuvent interpréter les informations de son en-tête ;

Time-Division Multiplex (TDM) Capable **:** interfaces qui commutent les paquets en fonction de la tranche de temps occupée dans un lien TDM ;

Lambda Switch Capable (LSC) **:** interfaces qui commutent les paquets en fonction de la longueur d'onde des données entrantes ;

Fiber Switch Capable (FSC) **:** interfaces qui commutent les paquets en fonction de leur emplacement physique (e.g., port, fibre).

Les LSPs dans GMPLS peuvent être créés seulement entre (et au travers) d'interfaces du même type. Un canal de contrôle est créé entre chaque paire de nœuds adjacents [27] grâce au Link Management Protocol (LMP) qui a la responsabilité de surveiller la santé des liens de données, gérer leur regroupement en liens TE et détecter les pannes de liens.

Parmi les extensions au protocole RSVP-TE, les plus importantes sont utilisées pour :

- représenter tous les types de commutation (voir les classes d'interfaces ci-haut) ;
- permettre aux LSPs d'exister entre des interfaces du même type (alors que ce n'est qu'entre deux routeurs pour MPLS) ;
- gérer les granulatités de bande passante pour chaque type de technologie de transport ;
- permettre l'existence de LSPs bidirectionnels ;
- permettre l'émission d'une notification rapidement suite à la détection d'une panne.

2.5.6 Propositions d'architectures combinant plusieurs éléments existants

Dans cette sous-section nous survolons des propositions d'architectures de réseaux futurs combinant des éléments déjà présentés précédemment.

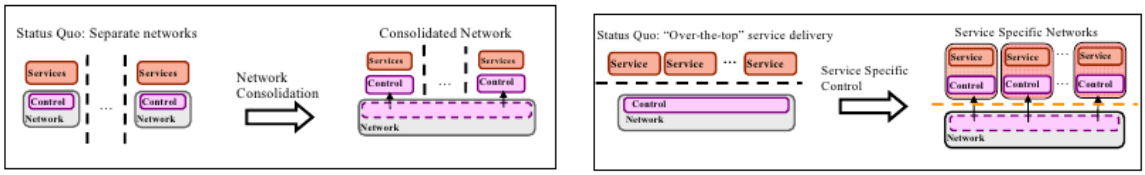
Les applications de la virtualisation des réseaux

L'approche proposée dans [24] est basée sur les concepts de virtualisation du projet 4WARD et réutilise donc les mêmes trois types d'éléments fondamentaux (voir la sous-section 2.5.2). Pour le plan de contrôle, les auteurs proposent d'améliorer GMPLS ainsi que les mécanismes définis par le groupe Path Computation Element (PCE) [28] de l'IETF. [29]

27. Le canal de contrôle peut ou non partager les liens physiques des canaux de données.
28. PCE permet de calculer des chemins de LSPs pour un plan de données (G)MPLS.
29. L'article ne donne toutefois aucun détail sur les modifications qui seraient nécessaires.

La virtualisation des réseaux permet deux importantes catégories d'applications :

1. le partage d'infrastructures qui présente deux variantes (Figure 2.8) :
   - la consolidation de réseaux dont le but est de combiner deux réseaux physiques existants sur une seule infrastructure physique ;
   - le *network slicing* qui permet de distribuer le contrôle des sous-réseaux partitionnés.
2. le contrôle combiné qui permet de gérer des domaines habituellement gérés séparément selon trois dimensions :
   - le contrôle de bout-en-bout dont le but est de regrouper le contrôle de plusieurs domaines en une seule entité commune ;
   - le contrôle multiniveaux qui permet de cacher la complexité du réseau de transport sous-jacent et optimiser les coûts de transport (le plan de contrôle unifié GMPLS ainsi que PCE constituent une première étape dans cette direction) ;
   - le contrôle de réseaux hétérogènes où différentes technologies d'accès sont gérées séparément mais qui pourraient être combinés grâce à la virtualisation.



(a) Consolidation de réseaux. (b) *Network Slicing*.

Figure 2.8 Variantes du partage d'infrastructures. *Source : [24]*

Partage d'infrastructures pour mieux résister aux pannes

Le principal sujet d'intérêt de [14] est le partage d'infrastructures dans le réseau d'accès afin d'augmenter la résistance aux pannes de liens, et ce sans augmenter significativement les coûts d'opération et les investissements de la part des MNOs. Les auteurs proposent d'utiliser une topologie en anneau plutôt qu'en étoile (ou en arbre) pour relier des groupes de eNodeBs au réseau cœur de chaque MNO. Une topologie en anneau a l'avantage d'offrir deux chemins pour joindre chaque MNO. Cependant, tous les liens entre les eNodeBs doivent être capables de supporter la charge totale de la grappe du réseau d'accès.

Une topologie en anneau suppose aussi qu'un eNodeBs puisse décider quel lien utiliser pour joindre un opérateur, avoir la capacité de détecter une panne et corriger sa table de routage. [30]

30. Les eNodeBs, tels que définis par 3GPP, n'ont pas à supporter la fonction de routeur IP.

Dans un autre ordre d'idée, si des liens micro-ondes relient les eNodeBs, la topologie en anneau a l'avantage de n'utiliser que quelques fréquences tout au plus, comparativement à un arbre.

Les auteurs proposent aussi de virtualiser les eNodeB, les relier avec des liens MPLS et utiliser OpenFlow/FlowVisor (sous-section 2.5.4) pour classifier les paquets, prioriser le trafic et déterminer ce qui est transmis vers le réseau d'un autre opérateur. Deux cas sont décrits :

- bande passante allouée statiquement, MPLS utilisé pour créer les LSPs statiques qui relient les eNodeBs ;
- bande passante allouée dynamiquement avec RSVP-TE pour la création des LSPs, utilisation d'un contrôleur centralisé pour gérer les demandes de réservations de ressources.

Partage d'infrastructures : réduire les coûts et conserver la flexibilité de gestion

On propose dans [54] une architecture NGN organisée en quatre plans :

- **Advanced Mobile Access (AMA)** : réseau d'accès radio composé d'eNodeBs, *femtocells*[31] et de stations de relai[32] ;
- **Optical Mobile Network (OMN)** : réseau d'agrégation comportant des commutateurs optiques reliés en anneaux (car les auteurs prévoient une augmentation importante du trafic P2P local) à des commutateurs de paquets pour les connexions externes ;
- **Virtual Mobile Network (VMN)** : réseau virtuel composé de OMNs physiques, permettant d'offrir des services globaux quel que soit l'endroit où se trouve l'usager ;
- **Service Delivery Network (SDN)** : réseau de services personalisés qui origine d'une évolution de IP Multimedia Subsystem (IMS).

Le partage d'infrastructures est abordé dans [37, 44] où l'on introduit un contrôleur centralisé permettant de séparer les opérateurs en réseaux virtuels. Le **Network Configuration Platform (NCP)** permet aux opérateurs de gérer eux-mêmes les ressources qui leur sont allouées (Figure 2.9). Le NCP gère les ressources physiques de façon grossière et les alloue aux opérateurs selon leurs besoins. Le but de l'architecture proposée est d'étendre le partage des infrastructures au-delà du RAN de façon à accomoder plusieurs MVNOs ayant des besoins spécifiques tout en garantissant l'isolation de ces derniers.

Le NCP permet de cacher aux MVNOs les détails des infrastructures physiques sous-jacentes. En effet, le rôle de ce dernier est de présenter aux MVNOs un plan de contrôle unifié alors qu'au niveau physique on peut retrouver GMPLS ou OpenFlow.

Chaque MVNO est connecté au NCP via leur **Virtual Network Controller (VNC)**. Le VNC est responsable de gérer les ressources virtuelles que le NCP leur a confiées. De plus,

31. Une *femtocell* est une petite station de base à usager domestique (pour 2–4 terminaux) ou pour petite entreprise (pour 8–16 terminaux). Permet d'augmenter la couverture à l'intérieur d'un bâtiment.

32. Une station de relai est reliée à un eNodeB (3GPP Rel. 10) afin de couvrir des endroits difficiles d'accès.

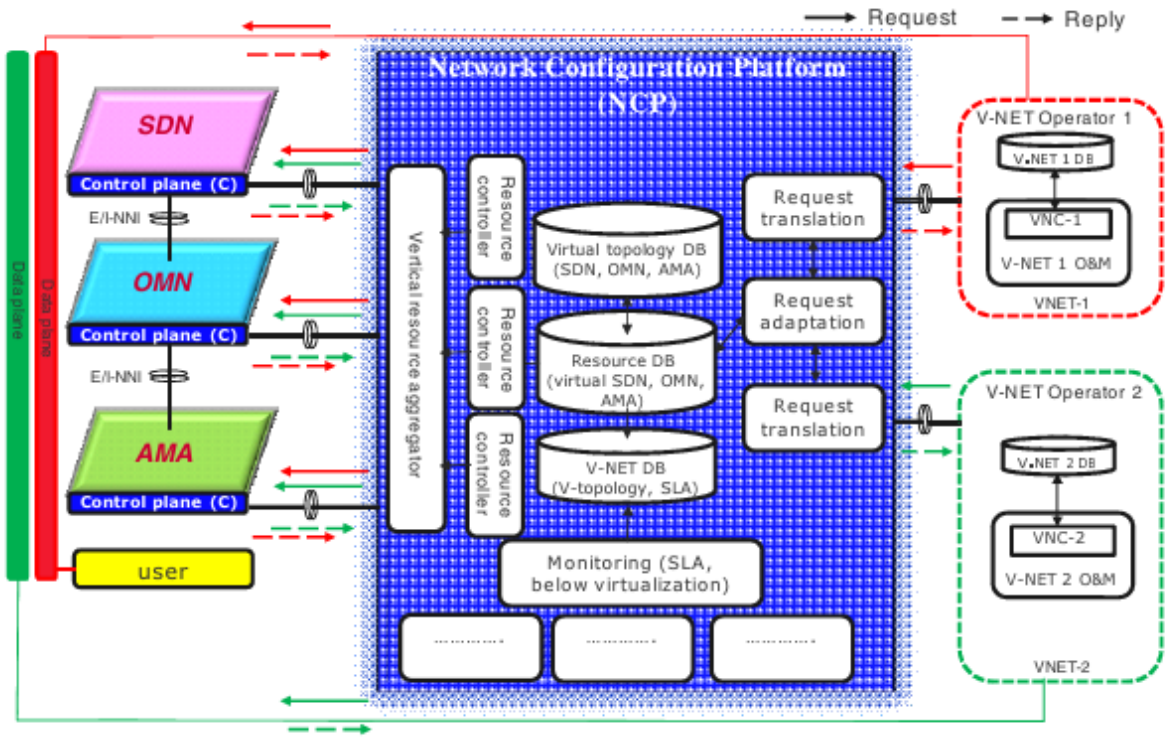


Figure 2.9 Network Configuration Platform (NCP). *Source : [37]*

le VNC émet des requêtes de réservations de ressources selon les besoins du MVNO.

L'approche proposée peut être comparée au concept *Infrastructure as a Service (IaaS)* associé au *Cloud Computing* [37]. En effet, les MVNOs jouent le rôle de clients qui peuvent adapter la configuration des infrastructures selon les besoins de leurs applications.

Partage d'infrastructures : approche basée sur le *Cloud Computing*

Le National Institute of Standards and Technology (NIST) des États-Unis propose la définition suivante de *Cloud Computing* [43] : *Cloud Computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computer resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.*

La discussion soulevée dans [17] vise à redéfinir ce qu'est un MNO découlant du paradigme de *Cloud Computing*. À titre d'exemple, les auteurs présentent un scénario dans lequel le RAN est virtualisé. L'avantage de cette approche est que les eNodeBs virtuels peuvent être programmés selon les besoins d'un opérateur, e.g., en configurant le RAN en tant que Single Frequency Network (SFN), en utilisant les canaux radio en Frequency-Division Duplexing

(FDD) ou Time-Division Duplexing (TDD), ou en programmant la taille des cellules de façon à servir adéquatement sa clientèle.

Bien entendu, seul le réseau cœur peut être implémenté dans un *cloud*.[33] Toutefois, les auteurs proposent d'intégrer la bande passante disponible au niveau physique aux ressources que l'on peut réserver d'une façon similaire aux autres ressources du *cloud*. Par contre, les problèmes d'optimisation de l'allocation de la bande passante partagée sont considérés hors sujet pour notre recherche. En conclusion, malgré que le vif de l'article soit hors sujet, l'application du paradigme de *Cloud Computing* aux réseaux cellulaires méritait d'être abordée.

2.5.7 Conclusions sur le partage d'infrastructures

De nombreux autres articles traitent de partage d'infrastructures mais furent écartés pour diverses raisons. En effet, nous n'avons pas considéré les articles qui présentent un modèle économique sans aborder son aspect technique, e.g., [12, 39, 42]. De plus, nous n'avons pas abordé le *roaming* qui est une forme actuelle de partage d'infrastructures qui permet aux usagers d'un opérateur de se connecter via le réseau d'accès d'un autre opérateur avec lequel ce dernier a un accord commercial. Une description des types de partage d'infrastructures est présentée dans [20, 42]. Les catégories de MVNOs sont présentées dans [17].

Il est clair que la tendance lourde en partage d'infrastructures pousse vers la virtualisation afin de permettre aux MVNOs de conserver un certain contrôle sur les infrastructures. En contrepartie, les premiers MVNOs ne possédaient qu'un système de facturation et un canal de distribution pour leurs applications.[34] Aujourd'hui, le défi de chaque MVNO consiste à se différencier du MNO et des autres MVNOs. Cela est possible grâce au *Cloud Computing*.

2.6 Revue d'articles sur la FMC

Dans cette section nous présenterons un ensemble de publications qui portent sur le sujet de la FMC. Le support de la FMC est important afin de décharger le réseau cellulaire lorsque c'est possible et offrir une couverture aux endroits mal desservis par ce dernier.[35]

Certaines publications traitant de FMC portent sur l'état des standards, l'optimisation des ressources radio ou sur le modèle économique favorisant le déploiement de la FMC (e.g., [18, 36]). Ces dernières ne seront pas considérées dans cette thèse car elles sont hors sujet.

33. Les serveurs d'applications peuvent être migrés dans le *cloud*. De même, certains nœuds du réseau cœur (e.g., HSS, P-GW) sont parallélisables. En conséquence, il serait possible de les dimensionner en fonction des besoins du MVNO, en ajoutant ou enlevant des machines virtuelles qui composent ces nœuds.

34. Ces MVNOs sont connus sous le nom de *Branded Resellers*.

35. Dans [18], l'auteur mentionne qu'environ 70% des appels 3G sont initiés de l'intérieur d'un immeuble. Le même taux est projeté pour la consommation de données.

L'importance (et la difficulté) de faire correspondre les classes de QoS d'un type d'accès à un autre est mentionnée dans [36]. La gestion de la QoS est aussi problématique lorsque l'on établit un tunnel sécurisé pour traverser le réseau d'accès d'un tiers (e.g., un AP Wi-Fi dans un café). En effet, il n'est pas possible pour le fournisseur d'accès de distinguer les flots de données à l'intérieur du tunnel encrypté et appliquer la QoS requise.

Dans [16], on s'intéresse à la mobilité et à l'utilisation simultanée des liens radio hétérogènes afin d'augmenter le débit global du terminal et gérer la micro-mobilité de ce dernier. Les auteurs proposent l'utilisation d'un serveur *proxy*[36] servant d'ancrage à la mobilité et à agréger les liens radio hétérogènes afin d'augmenter le débit du terminal. Le désavantage d'une telle approche est qu'elle n'est utilisable que dans le réseau d'attache de l'abonné. Par exemple, dans une situation de *roaming*, le point d'ancrage serait vraisemblablement situé dans le réseau d'attache de l'abonné (et donc éloigné de ce dernier), ce qui aurait pour effet d'amener le trafic au réseau d'attache et d'augmenter significativement le délai moyen. De plus, ce mécanisme empêche le routage normal des paquets de s'effectuer et rend difficile l'implémentation d'un *local breakout.*[37]

La FMC est parfois traitée conjointement au partage d'infrastructures. En effet, nous avons mentionné (voir la sous-section 2.5.6) que la virtualisation des réseaux permet l'émergence de deux importantes catégories d'applications, dont le *contrôle combiné* [24]. La dimension du contrôle combiné qui est liée à la FMC est le contrôle de réseaux hétérogènes. Ceci permet la création de stratégies connues sous le nom de « *Always Best Connected* » et au déploiement de mécanismes communs pour gérer les relèves.

L'intégration de la gestion de la QoS au niveau MAC pour les technologies Wi-Fi et WiMAX est proposée dans [22]. Les auteurs espèrent ainsi faire correspondre les mécanismes et les classes de services de ces deux technologies malgré le fait que la QoS sur Wi-Fi soit basée sur la différentiation de services et que celle sur WiMAX soit orientée connexion. Ils proposent un tableau de correspondances statiques entre les DiffServ Code Points (DSCPs), les services WiMAX et les classes de services 802.11. L'avantage de cette approche est sa simplicité. En contrepartie, le désavantage majeur de cette approche est qu'elle ne tient pas compte de la charge actuelle du réseau compte tenu que les associations sont statiques.

36. Un serveur *proxy* est un nœud intermédiaire par lequel transite le trafic des usagers, permettant une communication indirecte entre un client et un serveur distant. Un *proxy* peut servir d'ancrage à la mobilité, protéger le client (e.g., contre les virus informatiques) ou mémoriser des informations couramment accédées afin d'augmenter le débit.

37. Le *local breakout* est une optimisation en mobilité qui évite de faire transiter le trafic d'un usager par son réseau d'attache afin de réduire la congestion dans ce dernier et les délais des communications.

2.6.1 Support de la FMC pour l'accès 802.11

Pour le cas particulier de l'accès 802.11, un amendement à la couche MAC du standard actuel [28] traite spécifiquement des terminaux multimodes. En effet, l'amendement 802.11u [31] ajoute les mécanismes suivants :

- la découverte et la sélection d'un réseau externe ;
- les transferts d'informations provenant des réseaux externes ;
- support pour les alertes au niveau réseau et les communications d'urgence.

Un nouvel élément d'information (**Qos Map Set**) a été défini pour permettre à un AP de transmettre aux terminaux les associations entre les classes de services au niveau IP et les niveaux de priorité de la couche MAC. En particulier, cet élément associe des intervalles de valeurs des DSCPs aux 8 classes de priorité (sous-section 2.3.3) définies par la norme 802.1D [25]. Ces dernières sont ensuite directement associées aux ACs qui sont utilisées par le mécanisme EDCA afin d'accéder au canal radio.

2.6.2 Conclusions sur la FMC

Les principales motivations en faveur de la FMC sont que celle-ci peut décharger le réseau cellulaire des plus importants transferts de données et étendre la couverture à l'intérieur ou à des endroits où il n'est pas rentable d'ajouter une station de base additionnelle. De plus, considérant que beaucoup d'échanges de données sont réalisés à l'intérieur, la FMC permet d'augmenter le débit moyen du terminal, grâce à une connexion à haut débit à un AP local.

Toutefois, les considérations de sécurité (lorsque l'on utilise un réseau d'accès appartenant à un tier), de partage de la bande passante entre les usagers, de paradigme de QoS (différentiation de services *vs.* réservation de ressources) et d'association des classes de services constituent des problèmes qui ne sont pas entièrement résolus ou présentent des inconvénients encore aujourd'hui.

Enfin, il est important de mentionner que la FMC fait partie intégrante de la stratégie de l'industrie des communications mobiles qui est de préserver une connexion permanente entre l'usager et le réseau. Sans elle, plusieurs services envisagés n'exerceront pas l'influence escomptée auprès de la clientèle.

CHAPITRE 3

DÉMARCHES DE L'ENSEMBLE DU TRAVAIL DE RECHERCHE

Nous avons présenté au Chapitre 1 les objectifs principal et spécifiques de cette thèse. Nous verrons ci-après les liens qui existent entre les articles et les objectifs de recherche énoncés à la section 1.3.

Tout d'abord, le premier article dont le titre est « *A Potential Evolution of the Policy and Charging Control/QoS Architecture for the 3GPP IETF-based Evolved Packet Core* » identifie quatre tendances de fond actuelles au sein de l'industrie des communications mobiles :

1. le partage d'infrastructures physiques afin de réduire les coûts d'exploitation ;
2. la FMC afin d'offrir la meilleure couverture en termes de performances ou de coûts ;
3. la spécialisation des opérateurs dans le but de mieux cibler leur clientèle et ainsi se démarquer de leurs compétiteurs ;
4. la désintégration du modèle vertical d'opérateur avec pour objectif de se concentrer sur les activités de l'entreprise ayant la plus grande valeur ajoutée.

Dans un second temps, nous avons présenté deux architectures de gestion des politiques et de la QoS, soient l'architecture PCC de 3GPP et le RACS de TISPAN. L'analyse de ces architectures nous a permis de faire ressortir les points forts et les déficiences de chacune en ce qui concerne le partage d'infrastructures et le support de la FMC.

Ensuite, nous avons élaboré une liste de quatre concepts qu'une architecture de NGN devrait satisfaire afin de combler les besoins actuels et futurs des réseaux cœur :

1. la **séparation des services du réseau** afin d'assurer que les services soient accessibles, quel que soit l'accès utilisé ;
2. la **séparation des usagers du réseau**, ce qui facilite l'introduction d'un nouveau MVNO au sein d'un réseau cœur existant ;
3. la **séparation des rôles d'affaires** en domaines de QoS distincts facilitant le partage d'infrastructures entre les différentes entités administratives ;
4. un **réseau visité devrait pouvoir choisir le point d'attache** d'un terminal en situation d'itinérance car il est souvent le seul à connaître sa condition instantanée.

Les concepts présentés ci-haut sont à l'origine des définitions des trois rôles d'affaires qui constituent une partie de notre solution finale :

– le Access Network Provider (ANP) ;

– le IP Aggregation Network (IPAN) Provider ;
– le Network Service Provider (NSP).

Afin de respecter le concept de séparation des rôles d'affaires dans un réseau cœur basé sur l'architecture 3GPP, nous avons retiré du serveur PCRF la prise de décisions concernant les conditions du réseau et avons concentré ces fonctions dans un nouveau nœud. Le Network Policy Function (NPF) devient alors responsable de gérer les ressources de l'IPAN et d'y effectuer du contrôle d'admission. Ainsi, le IPAN peut devenir une entité administrative complètement indépendante des NSPs ; ces derniers sont maintenant tous considérés en tant que MVNOs qui se partagent un réseau d'agrégation IP.

Enfin, nous avons comparé notre solution aux architectures PCC de 3GPP et au RACS de TISPAN sur la base des concepts facilitant le partage d'infrastructures et le support de la FMC. Les résultats montrent que notre proposition, qui est une solution hybride tirant profit des points forts des architectures PCC du RACS, facilite grandement le partage d'infrastructures au niveau du réseau cœur en considérant tous les opérateurs en tant que MVNOs. De plus, notre architecture accomode très bien les opérateurs traditionnels puisque ces derniers n'ont qu'à jouer plus d'un rôle au sein de l'architecture globale.

Le second article intitulé « *A Network Policy Function Node for a Potential Evolution of the 3GPP Evolved Packet Core* » découle directement de notre premier travail et complète celui-ci. Les contributions significatives du second article sont de trois ordres :

1. nous proposons une amélioration à l'architecture présentée dans le premier article afin de simplifier cette dernière et réunifier les variantes GTP et PMIPv6 de l'EPC ;
2. nous analysons en détail les impacts de notre proposition sur chacun des nœuds de l'architecture PCC et faisons ressortir les nombreuses simplifications potentielles ;
3. nous décrivons le fonctionnement interne du NPF que nous avons proposé afin démontrer les avantages qui justifient le déploiement de notre architecture.

Dans un autre ordre d'idée, le mécanisme de QoS employé dans le EPS consiste à contrôler la QoS exclusivement à partir du réseau. [1] Toutefois, plusieurs applications émergentes ou de type P2P sont laissées pour compte puisqu'elles ne bénéficient d'aucun support de celui-ci.

Aussi, les technologies d'accès qui permettent le déploiement de la FMC peuvent implémenter l'un ou l'autre des paradigmes de QoS (réservation de ressources *vs.* différentiation de services) de même qu'un ensemble de paramètres de QoS qui leur sont propres. Cependant, 3GPP a normalisé des classes de services basées sur des garanties de performances absolues et défini un mécanisme de priorisation de l'allocation et de préservation des flots de données.

1. Jusqu'à la version 6 (inclusivement) de l'architecture 3GPP, seul le terminal pouvait signaler la QoS.

En dérivant les besoins en QoS des flots, pour chaque type d'accès, à partir du QCI et du Allocation and Retention Priority (ARP) autorisés par le PCRF, on peut plus facilement effectuer des relèves verticales tout en assurant adéquatement la QoS. Ceci constitue l'une des motivations à la base du troisième article intitulé « *A MultiAccess Resource ReSerVation Protocol (MARSVP) for the 3GPP Evolved Packet System.* »

En effet, MARSVP est un protocole de réservation de ressources, indépendant de la technologie d'accès, qui permet à un terminal de signifier au réseau ses besoins en QoS. De cette manière, on permet au réseau d'autoriser l'usage des ressources et d'initier les procédures de mise en place de la QoS ou, dans le cas des technologies d'accès nécessitant que le terminal initie la réservation de ressources, de spécifier au terminal les paramètres de QoS à utiliser.

Afin de valider le bon fonctionnement de MARSVP dans un environnement radio où les pertes de paquets sont inhérentes, une analyse formelle du protocole fut effectuée. Les pertes de paquets dues aux conditions radio furent simulées grâce à des arcs de transition parallèles entre une paire d'états. L'arc qui simule l'envoi d'un paquet qui ne sera jamais reçu ne provoque aucune synchronisation entre les automates des entités qui communiquent. Finalement, vingt propriétés du modèle ont été vérifiées afin de garantir le bon fonctionnement des machines à états du routeur d'accès, du terminal de même que l'application qui exécute sur le terminal.

En conclusion, les trois articles ont permis d'atteindre les objectifs secondaires de cette thèse. Collectivement, les articles ont contribué à l'amélioration de l'architecture PCC des réseaux mobiles 3GPP. En effet, l'évolution de l'architecture PCC simplifie grandement le partage d'infrastructures du réseau cœur et permet d'offrir un meilleur support de la FMC.

CHAPITRE 4

A POTENTIAL EVOLUTION OF THE POLICY AND CHARGING CONTROL/QOS ARCHITECTURE FOR THE 3GPP IETF-BASED EVOLVED PACKET CORE

Auteurs : Stéphane Ouellette, Laurent Marchand et Samuel Pierre.
Revue : *IEEE Communications Magazine*, vol. 49, no. 5, mai 2011, pp. 231–239.

Abstract

The 3rd Generation Partnership Project (3GPP) Release 8 defines the specifications of a low-latency, high-data rate all-IP Core Network (CN) capable of supporting real-time packet services over multiple access technologies, including the new Long-Term Evolution (LTE) access network.

The mobile communications industry matures. In order to remain competitive many operators now focus on reducing their capital and operational expenditures (CAPEX and OPEX), by outsourcing some of their business activities and/or by sharing some network infrastructures.

This paper presents a possible evolution of the 3GPP Policy and Charging Control (PCC) and Quality of Service (QoS) architecture to better support Fixed-Mobile Convergence (FMC) and more flexible CN sharing solutions. This goal is achieved with a clearer separation of the business roles (e.g., access and network providers) and the introduction of a Network Policy Function (NPF) for the management of the CN.

4.1 Introduction

The 3GPP Release 8 specifications define the Evolved Packet System (EPS) as an evolution of the General Packet Radio Service (GPRS). It features a flat Radio Access Network (RAN) architecture compared to a 3G packet-switched domain. As such, the functionalities of the Radio Network Controller (RNC) were split between the LTE base station (eNodeB), the Serving Gateway (S-GW) and the Mobility Management Entity (MME).

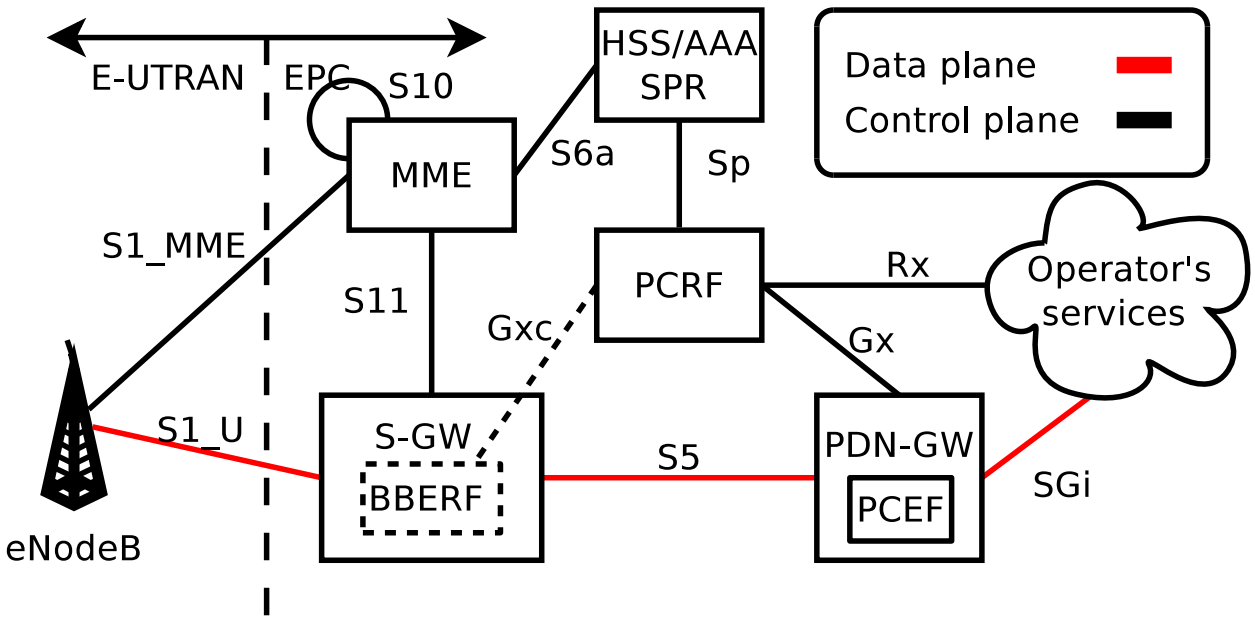


Figure 4.1 Main nodes of the EPS showing the LTE access (dotted elements are specific to [6])

As depicted in Figure 4.1, only eNodeBs are found in the RAN; all of the other nodes compose the Evolved Packet Core (EPC). Two network architecture solutions are defined for the EPC:

- The first one [5] is based on the GPRS Tunneling Protocol (GTP) and supports all 3GPP access technologies.
- The second solution [6] is based on the Internet Engineering Task Force (IETF) Proxy Mobile IPv6 (PMIPv6) protocol and defines some enhancements for non-3GPP access technologies.

This work derives from a research project that explores possible evolutions to the 3GPP network architecture beyond Release 8. Although most aspects of our proposal could be applied to the GTP-based architecture, we focused our efforts on the IETF-based solution.

The rest of this paper is organized as follows. Section 4.2 discusses the ongoing transformation of the communications industry. Section 4.3 describes the policy control and QoS architectures proposed by 3GPP and the Telecommunications and Internet-converged Services and Protocols for Advanced Networks (TISPAN) organization. Section 4.4 lists the requirements we believe should be fulfilled by the PCC and QoS architecture in order to better support FMC and more flexible CN sharing solutions. Section 4.5 details our proposal. Section 4.6 evaluates our solution based on the requirements listed in Section 4.4. Finally, Section 4.7 concludes and discusses the future works.

4.2 Changes in the industry

As the communications industry matures, many operators are more focusing at reducing costs rather than bringing innovation to the market [50]. In the future, the industry will be characterized by an increased specialization of the network operators (see Figure 4.2a) toward more specific market segments (e.g., teenagers, retired).

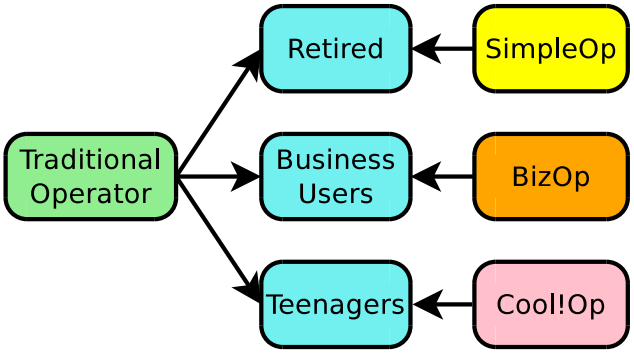
As part of this transformation, operators running both fixed and mobile accesses realized that they could further reduce costs by connecting their Access Network (AN) to a single, multi-access CN. Additionally, the emergence of multi-mode terminals allows the operators to provide their subscribers with either the best or the most cost effective available connection on all accesses.

The design of the EPC is in line with the rise of FMC which is a trend that aims to provide telephony and Internet access with a single device that can switch between local (e.g., Wi-Fi®) and wide-area (e.g., cellular) networks (Figure 4.2b). As there are many business advantages resulting from FMC, pure fixed or mobile-only operators are urged to build alliances to stay competitive.

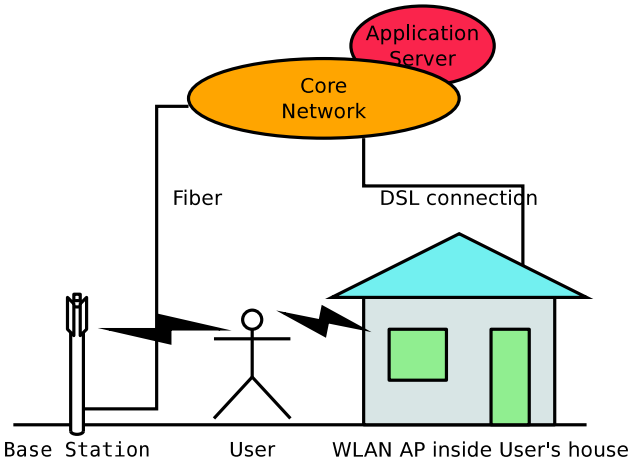
To reduce their costs, many operators restructure themselves by outsourcing parts of their network management activities or by sharing infrastructures with other operators [53]. The average revenue per user is the main driver for these reorganizations because it has remained constant in the past years, although voice and data traffic volumes per user are steadily growing.

Because of the fierce competition, the only way operators can increase their profits is to reduce their costs by focusing on "key assets" and "critical success factors" [20], i.e., their most profitable business activities or assets. As such, services are now being considered, beyond branding and coverage, as the prime differentiator between operators. Consequently, more advanced infrastructure sharing scenarios will emerge in the future.

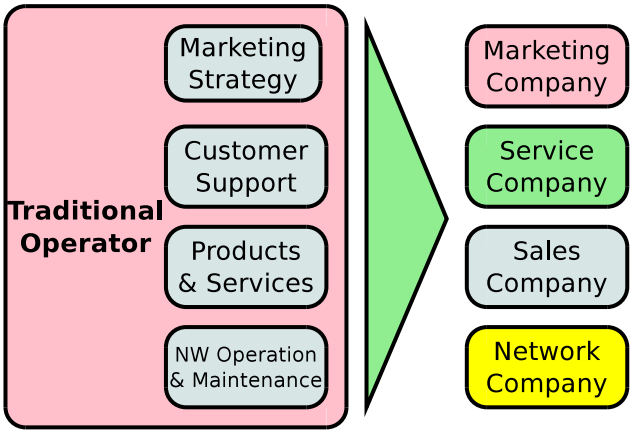
There are potential benefits to infrastructure sharing. First, a country's regulations can force operators to cover areas that are unattractive from a business point of view. Second, a "greenfield" operator could offer coverage to its customers in a previously uncovered region without having the burden to deploy a full-scale RAN; in return the operator already present gets the opportunity to leverage its RAN to generate more revenues [20]. Finally, because network dimensioning mostly depends on the peak demand, there are incentives for operators that have different subscriber bases (e.g., teenagers and business users that have disjoint traffic peaks) to share the RAN. This allows the average traffic level to increase without any significant increase in peak demand level because the traffic peaks occur at different times for the operators. Sharing infrastructures often leads to cost reductions.



(a) Specialization of operators.



(b) Fixed-Mobile Convergence.



(c) Disaggregation of the value chain.

Figure 4.2 Transformations of the communications industry

Based on this transformation of the industry, we foresee that it will progressively migrate from its vertical integration model to a disaggregated value chain (Figure 4.2c) in which the players focus on key business activities. Moreover, an increasing number of operators will offer services using several access technologies. These realities must be reflected by the logical CN architecture.

4.3 3GPP and TISPAN policy control/QoS architectures

This section presents the key features of the 3GPP PCC and TISPAN Resource and Admission Control Subsystem (RACS) architectures. TISPAN is a Next Generation Network (NGN) architecture focusing on fixed networks and Internet convergence. This section concludes with a list of issues regarding FMC and CN sharing support for both architectures.

4.3.1 The 3GPP PCC architecture

The PCC architecture [2] implements the policy and charging control functions of an EPC. For a detailed description of the PCC architecture, its procedures and QoS management, refer to [15, 47].

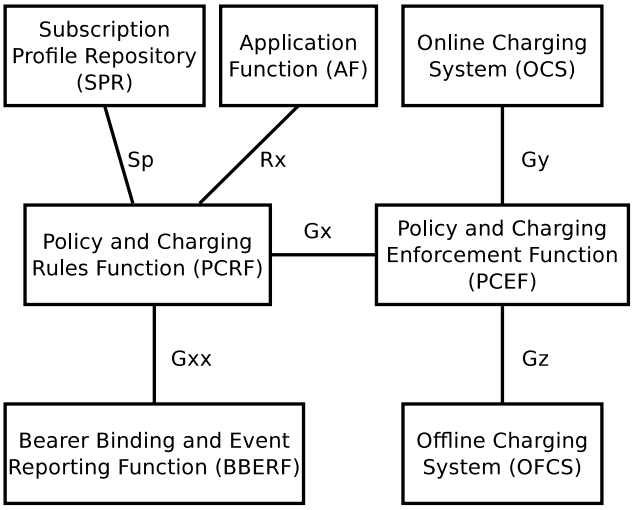
The Policy and Charging Rules Function (PCRF) decides how a Service Data Flow (SDF) shall be treated by the Policy and Charging Enforcement Function (PCEF). A SDF is a set of packet flows that matches a SDF template. A SDF template is a set of filters containing header parameters/ranges used to identify the packet flows constituting a SDF. The PCRF provides network control regarding SDF detection, QoS, gating and SDF-based charging, with the exception of credit management.

The PCRF takes policy decisions based on the session information received from the Application Function (AF) and on the user profile stored in the Subscription Profile Repository (SPR). The PCRF authorizes QoS resources defined by the QoS Class Indicator (QCI), Allocation and Retention Priority (ARP), guaranteed/maximum bitrates (GBR, MBR). When an AF request is granted, the PCRF sends a QoS rule to the Bearer Binding and Event Reporting Function (BBERF) and a PCC rule (basically a QoS rule with gating and charging informations) to the PCEF.

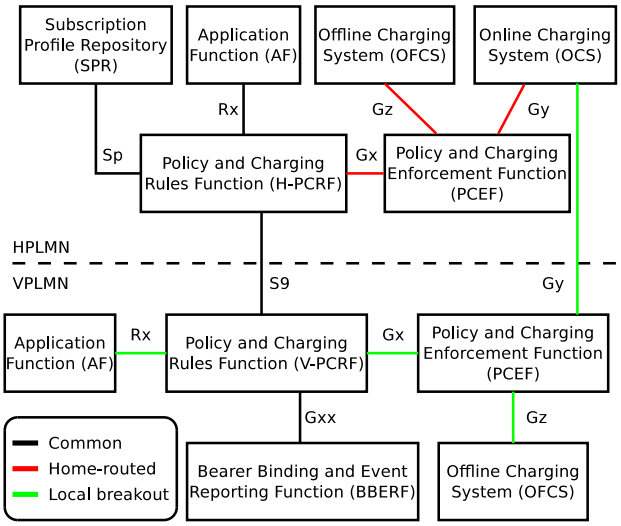
The PCEF encompasses SDF detection and measurement, gate and QoS enforcement, online and offline SDF-based charging functionalities and event reporting to the PCRF. The PCEF is implemented into the Packet Data Network (PDN) Gateway (P-GW) [1].

1. The PCEF can also be implemented into the evolved Packet Data Gateway (ePDG) but untrusted non-3GPP networks are out of scope.

The BBERF is implemented into the Access Edge Gateway (AEG) for each type of access. Its main tasks are event reporting to the PCRF and bearer binding [2]. The BBERF exists because IETF mobility protocols focus on routing only and carry no information about the bearers' QoS properties.



(a) *Non-roaming* architecture.



(b) *Home-routed* and *local breakout* roaming cases.

Figure 4.3 Logical PCC architecture for the IETF-based EPS (*roaming* and *non-roaming* cases)

2. A bearer is a logical IP data path between the User Equipment (UE) and the network with specific QoS properties. Bearer binding is the association between a SDF and the IP-CAN bearer (also known as a PDP context for GPRS).

Mobility is supported within 3GPP/3GPP2 accesses but session continuity that involves non-3GPP accesses is still under development. The non-roaming architecture is presented in Figure 4.3a while Figure 4.3b illustrates two roaming cases:

– The *home-routed* case forces all user traffic to be tunneled back to the home network.

– The *local breakout* case allows the UE to be connected to a P-GW in the visited network.

Infrastructure sharing is possible in both the RAN and parts of the CN. RAN sharing involves not only sharing the RAN nodes but also frequency pooling. Two basic scenarios for infrastructure sharing are defined for the EPC [3]:

– Multi-Operator Core Network (MOCN) in which only the RAN nodes are shared.

– Gateway Core Network (GWCN) configuration where RAN and S-GW are shared.

4.3.2 The TISPAN RACS

The RACS [52] offers to AFs a mechanism to reserve resources from the network. Figure 4.4 shows the logical architecture of the RACS (only a subset of the interfaces are visible and the charging aspect is left out of scope) and of the user plane nodes. Refer to [51] for an overview of the TISPAN Release 2 architecture.

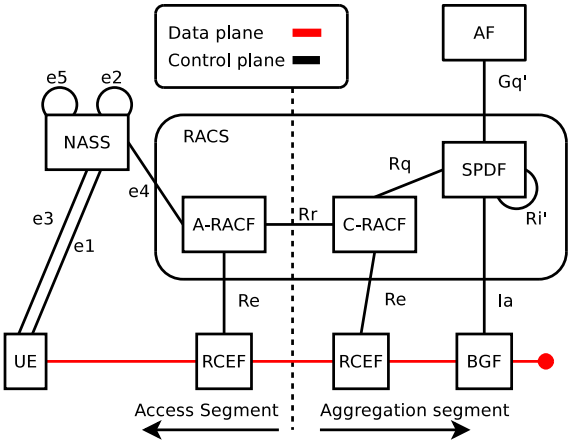


Figure 4.4 TISPAN RACS logical architecture

The Resource Control Enforcement Function (RCEF) supports a number of elementary functions: opening and closing gates, packet marking, policing of incoming traffic, resource allocation for upstream and downstream traffic.

The Border Gateway Function (BGF) interfaces two IP domains. It encompasses the functionality of the RCEF and also supports additional elementary functions: usage metering, Network Address Translation (NAT), hosted NAT traversal, etc. The TISPAN BGF is very similar to a 3GPP PCEF.

The Network Attachment SubSystem (NASS) is responsible of authentication and authorization based on the user identity, IP address allocation and configuration of the user's device via the e1 reference point. In case the user device is a Customer Network Gateway (CNG), the e3 reference point is used to configure it. At last, the NASS supports roaming with the help of two reference points:

- The e5 interface is used to proxy user authentication requests to the home network.
- The e2 reference point enables AFs to retrieve information about the characteristics of the IP-connectivity session used to access such applications (e.g., network location information) from the Connectivity session Location and repository Function (CLF), a sub-component of the NASS.

The RACS itself is composed of two primary function blocks:

- The Access RACF (A-RACF) is a functional entity that manages resources in the AN and performs admission control, taking into account the user's access profile retrieved from the NASS through the e4 reference point. A Core RACF (C-RACF) is also defined to manage the aggregation network's resources (but it is user-unaware).
- The Service Policy Decision Function (SPDF) is a functional entity acting as a policy decision point for service requests received from an AF. It applies operator-defined policy rules that specify a service's resource needs, NAT and firewall traversal rules, etc. It doesn't consider the user identity as it has no direct access to the NASS. The SPDF performs a coordination function between the AF, A-RACF and BGF. It also supports charging. Roaming is supported over the Ri' interface that links the SPDF in the home network to the one in the visited network. As of TISPAN Release 2, only nomadism is supported, not mobility.

4.3.3 FMC and CN sharing issues in 3GPP and TISPAN networks

When one considers the transformation of the communications industry as seen in Section 4.2, some issues appear for both architectures.

Suppose that two operators share a 3GPP CN. Because the operators want to offer their own applications to their users, both need a policy server that takes service requests and user profiles as inputs to the policy decision process, independently of the CN conditions. In addition to taking user- and service-related policy decisions, the 3GPP PCRF directly controls the policy enforcement points (PCEF, BBERF), maps QCIs to DiffServ Code Points (DSCPs) and determines the charging rules. Clearly, two operators can't control the network resources independently nor decide the network policies to enforce.

On the other hand, the TISPAN architecture is handicapped regarding global mobility, FMC and roaming because the user is tied to the AN. This also makes it more difficult to integrate Mobile Virtual Network Operators (MVNOs) [3].

4.4 Requirements on the Policy Control and QoS architecture to support CN sharing/FMC

A number of concepts must be respected to ensure that a PCC/QoS architecture meets today's and future needs of the CN. These concepts shall maximize an operator's flexibility to follow any evolution path desired, such as broadening the palette of accesses that a user can choose from, or growing into elaborate network sharing scenarios to reduce costs. FMC and flexible CN sharing have no relationship *a priori* but an evolution of the 3GPP architecture that satisfies the requirements below will improve both aspects. The following four requirements consider FMC and infrastructure sharing in general.

First, the network architecture must clearly separate the services from the network. This implies a separation of policy control applied to the service requested and the user's profile on one side, from the network policies [4] and resource management on the other side. Therefore, this approach would allow service convergence to take place because the users can invoke the services they subscribed to from any AN. Moreover, potential service duplication is avoided because many services are common to all accesses.

Second, the PCC and QoS architecture must separate the subscriber management (service and user policies, IP address allocation, authentication) from the network management. As a consequence, this separation facilitates the integration of a MVNO on top of an existing CN.

The third requirement stipulates that each business/network entity should directly control its assets, especially when they are shared between multiple "clients". For example, a CN operator sharing its assets between two Network Service Providers (NSPs) must ensure that the Service Level Agreements (SLAs) are respected. As a result, the NSPs cannot directly reserve the resources of the CN; they must send a request to a resource manager in the CN for prior approval.

The last requirement applies to FMC support of roaming users. The visited network should be involved in AEG selection because the home network is rarely aware of the visited network's conditions. Also, the visited network should be able to move the UE between accesses based on the evolution of the visited network's conditions.

3. A MVNO has no network infrastructures nor spectrum license but it has a subscriber base and can offer applications to its subscribers.

4. An example of network policy could be to keep a minimum of the total bandwidth for best-effort traffic.

4.5 A potential solution

First, we define the following business roles [5] to support network sharing scenarios:

- The Network Service Provider (NSP) hosts AFs (IMS [6] and non-IMS) and bills the user (by volume, duration, QoS requested, etc). It defines user- and service-related policies (hosts a PCRF), owns the subscriber base, authenticates the user (hosts a HSS) and controls the BGFs in the P-GW. It interacts with other NSPs for roaming.
- The IP Aggregation Network (IPAN) provider sets up SLAs with the NSPs to which it offers transport services. It owns the edge nodes (e.g., S-GW, P-GW) and interacts with the underneath transport architecture. It links Access Network Providers (ANPs) to the NSPs and offers infrastructure services (e.g., antivirus DPI [7] engine, traffic localization) to the supported NSPs.
- The Access Network Provider (ANP) manages the access network resources according to a set of local policies and enforces the SLAs made with the IPAN operator(s).

Second, this paper builds on top of [38] which has previously described possible enhancements to the mobility architecture. For roaming users it introduced a local mobility anchor in the visited P-GW to clearly separate local and global mobility. This anchor allows the visited network to play a role in the selection of the AEG because it can track the network conditions and move the UE between accesses. The proposed roaming architecture is illustrated in Figure 4.5a.

Figure 4.5b shows a hypothetical network's control plane that emphasizes on the separation of the business roles and involves two NSPs and two ANPs. The data plane is similar to the one from Figure 4.1, except that the IPAN can offer some infrastructure services to the NSPs. Their location is intentionally unspecified as some (notably the location-based) services should be implemented close to the UE to maximize their efficiency.

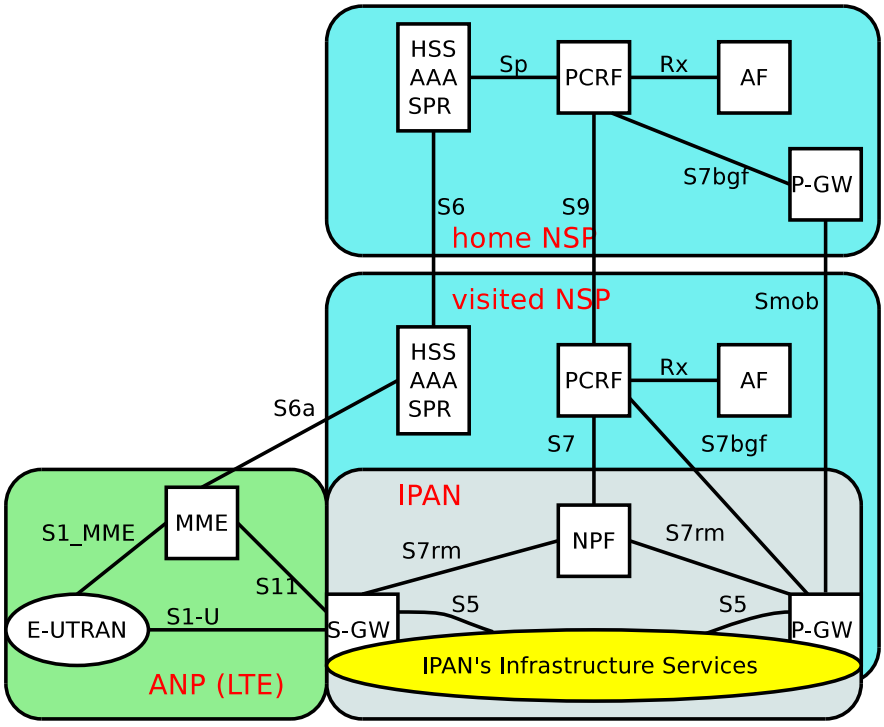
Infrastructure services can be considered as advanced value-added functions that are shared by numerous applications. They are a generalization of the concept of BGF (which are only implemented into a P-GW) and can benefit both IMS and non-IMS applications.

Each NSP hosts a P-GW that belongs to the IPAN but yields some of its border gateway functions to the NSP. The PCEF functions under IPAN control are QoS enforcement and gating while the others are under NSP control.

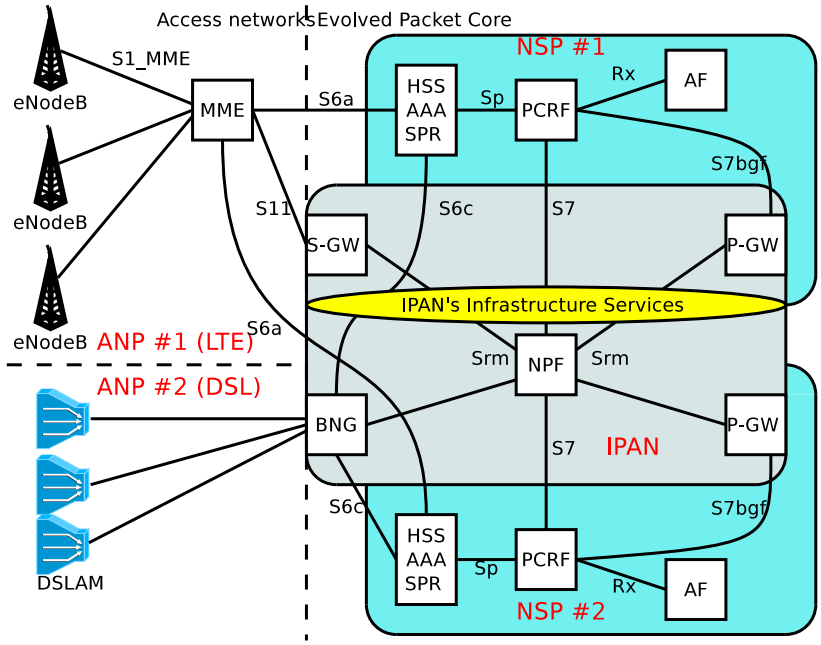
5. Note that traditional operators play all roles.

6. The IP Multimedia Subsystem (IMS) is an architectural framework for delivering IP multimedia services.

7. Deep Packet Inspection (DPI) implies that the protocol headers as well as the data traffic are inspected.



(a) Roaming architecture valid for both *home-routed* and *local breakout* cases (control and data planes).



(b) *Non-roaming* architecture featuring CN sharing and multiple access networks (control plane only).

Figure 4.5 Proposed evolution of the 3GPP PCC and QoS architecture

The Network Policy Function (NPF) was introduced to manage the IPAN resources and the infrastructure services it offers to the NSPs. Its main purpose is to completely separate the underlying transport network from the NSPs. As a result, NSPs are considered as MVNOs. The NPF enforces the SLAs between itself, the ANPs and the NSPs.

The NPF applies network policies [8] according to the IPAN conditions regardless of the user or service requested. It can modify the network policies based on the situation or time of the day (normal, overload, emergency situation).

The NPF is access-agnostic to the PCRF but relies on access-specific mechanisms to perform QoS and admission control in the AN. As such, the NPF is a mediator between the 3GPP CN and the policy infrastructure that might exist for any given access technology.

For scalability reasons the NPF only performs coarse-grained resource management based on pre-congestion notifications received from the AEGs or P-GWs. This allows the NPF to assist the UE during inter-AEG handovers, by telling the UE for example not to switch to the new AEG because the latter notified the NPF of an imminent congestion situation. At least two options for pre-congestion notifications must be studied:

- If the IPAN is a plain DiffServ domain, marking each packet with congestion informations may be appropriate.
- If the IPAN operates on top of virtual circuits (e.g., IP-MPLS, MPLS-TP, PBB-TE), the tunnel ingress node can easily determine when a tunnel is about to become congested.

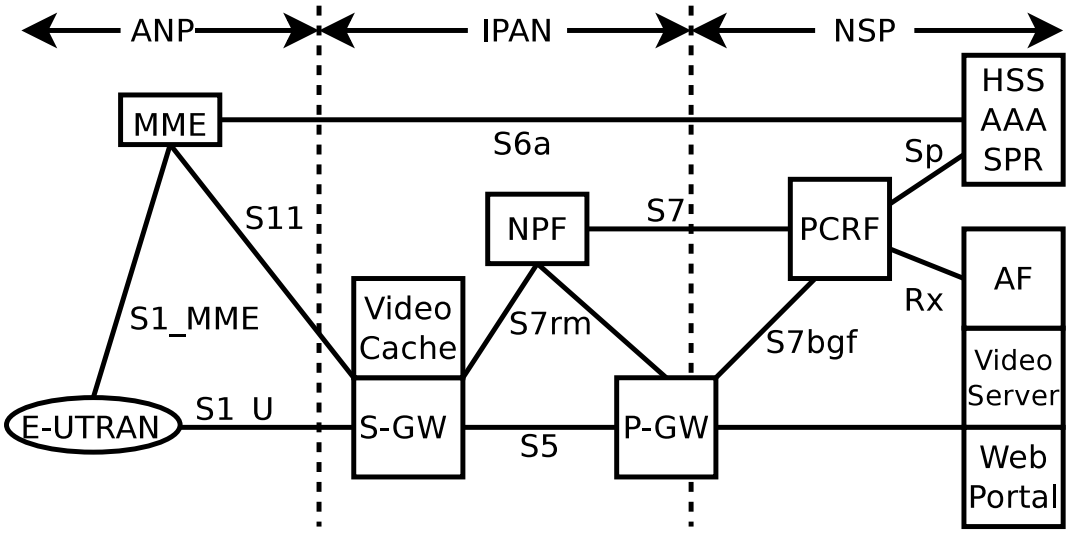
In both cases, when the ingress node experiments congestion, the NPF is notified so that it will reject any new resource reservation for the congested GBR QoS class until further notice.

At last, the NPF maps the 3GPP QCIs to L3 QoS classes (DSCP codes) and coordinates the mappings if multiple QoS domains exist in the IPAN. The NPF supports Priority Services based on the ARP of the resource reservation.

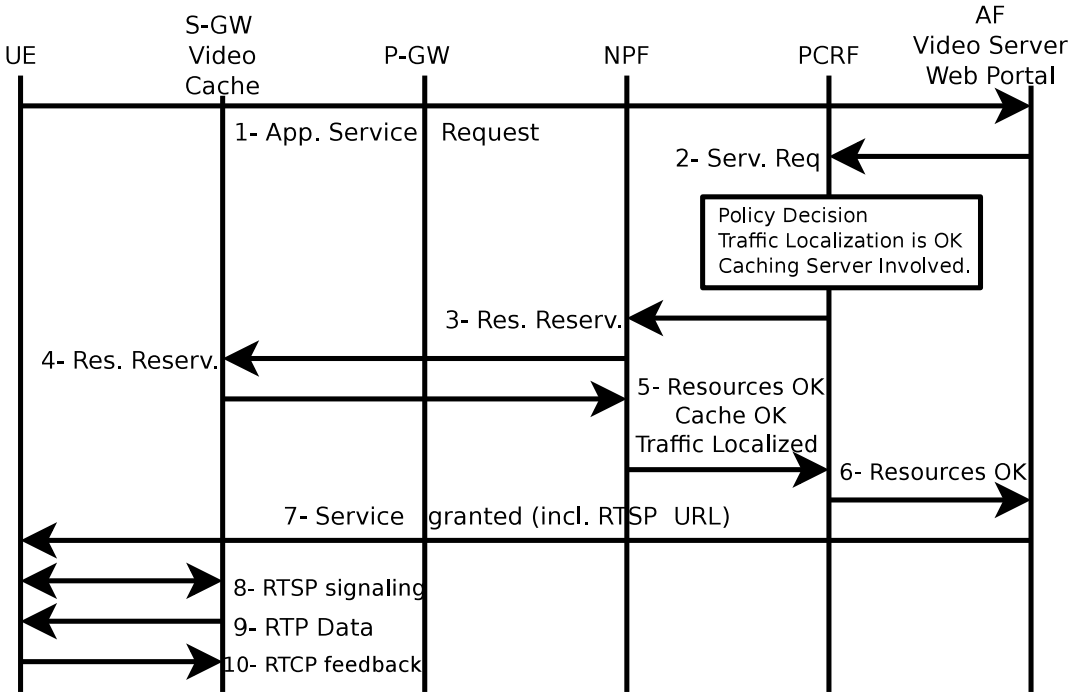
Three Diameter-based reference points are defined in our solution:

- S7 carries the SDF QoS parameters and supports infrastructure service invocation. S7 is based on Gxx and will include the yet to be defined Attribute-Value Pairs (AVPs) to announce and invoke infrastructure services.
- S7bgf controls the border gateway functions and specifies the charging rules to the P-GW for each SDF. S7bgf is based on Gx but doesn't specify the SDF QoS parameters.
- Srm carries the QoS rules of each SDF to the AEGs and P-GWs for QoS rate enforcement and QCI mapping. Srm is based on Gxx.

8. For example, a network policy could be that the sum of bandwidths of some L3 QoS classes must not exceed a given total because they are mapped to a single L2 class.



(a) Architecture for VOD and caching example.



(b) Sequence diagram for this example.

Figure 4.6 Infrastructure service example featuring video caching collocated with the S-GW

Figure 4.6 depicts an example of infrastructure service in which a Video On-Demand (VOD) server uses a cache (collocated with the S-GW) in order to store the most popular movies. This example also features Traffic Localization which prevents user traffic to be tunneled back to the P-GW unless required. Step 1 shows the application service request sent by the UE to the AF. Step 2 is the service request sent by the AF to the PCRF, describing the resource needs and infrastructure services to invoke. The PCRF decides that the service is entitled to use Traffic Localization and the video cache, then sends a resource reservation request to the NPF (Step 3). The NPF takes a network policy decision then sends the resource reservation request to the S-GW in order to setup the resources in the AN and bind the caching server to the corresponding GTP tunnel (Step 4). A positive answer is returned to the AF (Steps 5 and 6). In Step 7, the AF returns a Real-Time Streaming Protocol (RTSP) Universal Resource Locator (URL) to the UE that actually points to the S-GW. At last, the UE connects to the video caching server and exchanges Real-Time Protocol (RTP) traffic and Real-Time Control Protocol (RTCP) feedback.

4.6 Discussion

This section discusses how the requirements from Section 4.4 are met by TISPAN, 3GPP and our proposal. Table 4.1 summarizes the results.

4.6.1 Separation of AN from services

From the AF's point of view, all three architectures fully separate the services from the AN. The architectures are access-agnostic and use generic QoS parameters (QCIs are used for 3GPP and our proposed evolution) to specify the packet forwarding treatment. Our proposal and 3GPP use the Rx reference point to link an AF to the PCRF. Similarly, this functionality is achieved by TISPAN with the Gq' reference point that links an AF to the RACS.

4.6.2 Separation of user management from network management

The 3GPP PCC architecture and the evolution we propose feature a Home Subscriber Server (HSS) to store subscriber credentials and service profiles. Both own a PCRF that takes user- and service-based policy decisions, for roaming and non-roaming cases, regardless of the network policies and resource management. However, our solution explicitly defines the NPF to take network policy decisions and perform high-level resource management. In the 3GPP model, static network policies can be defined into the P-GW but this depends on node configuration.

Table 4.1 Brief comparison of PCC and QoS architectures: 3GPP, TISPAN and our solution

Concept	TISPAN	3GPP	Suggested evolved 3GPP PCC architecture
Separation of services from the AN.	Full (Gq' interface between the RACS and AF).	Full (Rx interface between the PCRF and AF).	Full (Rx interface between the PCRF and AF).
Separation of user and network management.	AN authenticates user, allocates IP address. A-RACF takes user policy decisions.	User auth., IP addr. alloc. and user-/service-based policy decisions done by CN. No clear separation with IP transport.	User auth., IP addr. alloc. and user-/service-based policy decisions done by CN. Network mngt. done by the NPF in the IPAN.
Separation of the business roles. Distinct domains for policy and resource mngt.	Separation of aggregation and AN resource mngt. and policy control. No distinction between IP transport and NSP.	Distinguishes RAN and CN operators. CN sharing not standardized beyond S-GW.	Clear separation of NSP, ANP and IPAN provider business roles. Flexible network sharing solution.
Local mobility anchor in visited network.	N/A.	In S-GW for home-routed scenario.	In the P-GW in all cases.

On the other hand, TISPAN doesn't meet this requirement because the AN is user-aware. The NASS retrieves the subscriber profile from the Authentication, Authorization and Accounting (AAA) server then forwards it to the A-RACF over the e4 interface. Note that the TISPAN network architecture presents serious handicaps for FMC, roaming and eventually for mobility support. As a result, the A-RACF admission control is based on resource availability, AN policies and on the subscriber profile.

4.6.3 Separation of business roles into independent PCC/QoS domains

The 3GPP architecture partially meets this requirement because Universal Mobile Telecommunications System (UMTS) was originally built under the principle "one operator, one radio access network" [1] and as a result the standard lacks some functionalities that facilitate network sharing. In fact, the only business roles identified by 3GPP are RAN and CN (grouping

the NSP and IPAN roles) operators. CN node sharing isn't standardized beyond the S-GW.

TISPAN partially meets this requirement because it considers the access and core networks as two distinct domains. Additionally, it defines the C-RACF which provides a separation of resource management between the access and aggregation networks. However, as with 3GPP, there is no distinction between the entity that provides IP transport and the one that owns the user database and offers application services.

Our proposal specifically introduces the NPF to manage the IPAN and share the CN resources in multiple NSP scenarios. Each AN is responsible for its own resource management but the NPF relies on access-specific mechanisms to determine if there are available resources in both the aggregation and AN. It is worth noting that most access technologies don't define a central AN policy server and resource manager.

4.6.4 AEG selection when roaming

This requirement is not met in the 3GPP Release 8 specifications for the *home-routed* case. The UE's traffic is tunneled between the visited S-GW and the home P-GW. The visited network cannot move the UE to another AEG. This does not apply to TISPAN because only nomadism is supported in Release 2. At last, our solution inherited a local mobility anchor in the *home-routed* and *local breakout* cases.

4.7 Conclusion

This paper suggested possible enhancements to the 3GPP PCC architecture to ease the design of advanced infrastructure sharing scenarios and better support FMC. These are built on top of the following concepts:

- Separation of the services from the AN that is essential to full service convergence.
- Separation of user/network management to simplify the roaming support and facilitate the creation of MVNOs.
- Separation of business roles into independent PCC/QoS domains that simplifies the design of complex sharing scenarios that not only cover networks that were designed with sharing in mind but also those that introduce it later on as an evolution path that follows a NSP's change of business strategy.
- Local mobility anchor in visited network for roaming scenarios. The UE can move to another AEG based on network conditions.

The next steps of this work are to fully specify the reference points and network nodes affected. Two additional work items are the support of infrastructure services and inter-AEG handovers.

CHAPITRE 5

A NETWORK POLICY FUNCTION NODE FOR A POTENTIAL EVOLUTION OF THE 3GPP EVOLVED PACKET CORE

Auteurs : Stéphane Ouellette, Laurent Marchand et Samuel Pierre.
Revue : Accepté moyennant des corrections mineures dans le journal *Elsevier Computer Networks* en août 2012.

Abstract

The 3rd Generation Partnership Project (3GPP) Release 8 architecture defines the specifications of the Evolved Packet Core (EPC), which is an Internet Protocol (IP)-based multiaccess Core Network (CN) capable of supporting high-speed real-time packet services, and a new radio access widely known as Long-Term Evolution (LTE).

In a previous paper we studied an evolution of the Policy and Charging Control (PCC) and Quality of Service (QoS) architecture to better support both Fixed-Mobile Convergence (FMC) and CN sharing solutions. Our solution introduced a Network Policy Function (NPF), a centralized network policy decision point, to manage the CN resources and act like a mediator when many Network Service Providers (NSPs) and/or Access Network Providers (ANPs) are involved.

The NPF hides the details of the underneath network topology from the NSPs which operate over virtualized resources. This paper focuses on the features of the NPF as well as the interactions between itself and the other CN nodes.

5.1 Introduction

The 3GPP Release 8 architecture is referred to as the Evolved Packet System (EPS). The EPC is an evolution of the General Packet Radio Service (GPRS) that contains no circuit-switched nodes, as opposed to the 3G CN. The Evolved Universal Terrestrial Radio Access Network (E-UTRAN) is built exclusively of LTE base stations (eNodeBs).

3GPP proposes two slightly different network architectures for the EPC (see Figure 4.1):

- the first proposal [5] uses the GPRS Tunneling Protocol (GTP) for QoS, bearer and mobility management. It supports all 3GPP technologies (GSM, UMTS and LTE);
- the second solution [6] integrates several architecture enhancements for some non-3GPP accesses (e.g., HRPD, WiMAX and DSL). The Internet Engineering Task Force (IETF)

Proxy Mobile IPv6 (PMIPv6) drives network-based mobility. The Gxc reference point manages QoS and bearer signaling.

The PCC architecture [2] is similar for both EPC variants but this is transparent to the User Equipment (UE). PCC allows service and user differentiation and also determines the charging method to apply. Service differentiation ensures that the user's Quality of Experience (QoE) for each running service is always acceptable, as over-provisioning radio resources is expensive. User differentiation allows an operator to apply differentiated treatments to a given user's traffic based on the type of subscription the user has. Finally, charging can be online/offline, by traffic volume and/or duration, event-driven, etc.

This work derives from a research project that explores possible evolutions to the 3GPP PCC network architecture beyond Release 8. Even though most aspects of our solution could be applied to the GTP-based architecture, we focused our efforts on the IETF-based solution.

This paper is organized as follows. Section 5.2 introduces the PCC architecture. Section 5.3 summarizes the ongoing transformations of the mobile communications industry. Section 5.4 explains why an evolved PCC architecture is needed and provides the foundations of our work. Section 5.5 describes the impacts of our works onto the existing 3GPP nodes. Section 5.6 further details the NPF functions and processes. Finally, Section 5.7 concludes and discusses future works.

5.2 The 3GPP PCC Architecture

The policy engine of the PCC architecture (see Figure 4.3) is the Policy and Charging Rules Function (PCRF). The PCRF derives PCC rules from the information below (refer to [15, 47] for details):

- the QoS required by the Service Data Flows (SDFs) of the Application Function (AF);
- the subscriber profile (e.g., corporate *vs.* private, bronze/silver/gold subscription, roaming *vs.* non-roaming) retrieved from the Subscription Profile Repository (SPR);
- operator-defined policies for some services;
- the actual load of the transport network;
- the current time and date;
- the user's current Radio Access Type (RAT) in case of a multiaccess EPS.

From the inputs above, PCC rules determine:

– the Guaranteed and Maximum Bit Rates (GBR, MBR), if applicable;

– the SDFs QoS Class Indicator (QCI) and Allocation and Retention Priority (ARP);

– the appropriate charging method;

– SDF templates (i.e., packet flow filters).

The function which enforces policy rules is the Policy and Charging Enforcement Function (PCEF). The PCEF provides SDF detection and measurement, gate and QoS enforcement, online/offline SDF-based charging and event reporting to the PCRF. The PCEF is part of the Packet Data Network (PDN) Gateway (P-GW). [1]

As opposed to GTP, which manages QoS and mobility, PMIPv6 carries no information about a bearer's QoS properties. As a consequence, the IETF-based PCC architecture features an additional function that is implemented in the Access Edge Gateway (AEG), e.g., the Serving Gateway (S-GW), for each type of access. This logical node is known as the Bearer Binding and Event Reporting Function (BBERF). [2] Following a policy decision, the PCRF sends the PCC rule to the PCEF and a QoS rule (i.e., a PCC rule without charging information) to the BBERF.

The charging functions of PCC are provided by two systems that interact with the PCEF and the operator's billing system:

– the Offline Charging System (OFCS) which processes Charging Data Records (CDRs) from various network elements only after the usage of the network resources is complete;

– the Online Charging System (OCS) which performs functions similar to the OFCS but requires additional handling for authorizing network resources prior to their usage.

Application Functions (AFs) transfer dynamic session information to the PCRF on the Rx interface and are notified about bearer level events.

Finally, the non-roaming PCC architecture is depicted in Figure 4.3a while Figure 4.3b illustrates the two supported roaming cases:

– the *home-routed* case forces all user traffic to be tunneled back to the home network;

– the *local breakout* case allows the UE to be connected to a P-GW in the visited network.

5.3 Changes in the Industry

The mobile communications industry matures; services have become the prime differentiator between operators, beyond branding or coverage. At the same time, some operators are

1. A PCEF can also be implemented within an evolved Packet Data Gateway (ePDG) but untrusted non-3GPP networks are out of scope.

2. Bearer binding (i.e., mapping SDFs to IP tunnels with adequate QoS) in [5] is done by the Bearer Binding Function (BBF) which is part of the PCEF.

becoming increasingly specialized by targeting more specific market segments (e.g., teenagers, retired). As a consequence, many operators now focus on their key business activities and seek to reduce their costs by outsourcing some of their less profitable activities or by sharing infrastructures with some of their competitors [20, 50, 53].

Today's terminals can communicate using several access technologies. FMC is a trend aiming to provide Internet access and telephony on a single device capable of switching between local- and wide-area networks. Using FMC, operators running both fixed and mobile accesses can offer their users either the best or the most cost effective available connection on all accesses. These operators can further reduce their costs by connecting their access networks to a shared CN.

From this evolution of the industry, we foresee that it will progressively migrate from its vertical integration model to a disaggregated value chain in which the players focus on their key business activities. Also, an increasing number of operators will offer services over several access technologies. These realities must be reflected by the logical CN architecture [46].

5.4 A Potential PCC Evolution

As stated in Section 5.3, the mobile communications industry is evolving. As a result, the PCC architecture must be able to meet today's and future needs of the CN. In this section, we present the foundations of our work. First, we highlight the FMC and CN sharing issues for two different policy architectures. Then, we list the requirements that the PCC architecture must meet to better support FMC and CN sharing. At last, we conclude this section by providing an overview of our solution.

5.4.1 FMC and CN Sharing Issues in 3GPP and TISPAN Networks

The 3GPP PCC architecture has been compared to the TISPAN [3] Resource and Admission Control Subsystem (RACS) [52] regarding FMC and CN sharing support. The conclusions of this evaluation [46] are summarized below.

The 3GPP PCC architecture was built under the principle that a CN belongs to a single operator [1]. In a hypothetical shared CN scenario, each operator would host a PCRF to take policy decisions. Obviously, two operators cannot simultaneously control the CN resources nor decide which network policies to enforce. Furthermore, dynamic resource sharing (i.e., allowing an operator to borrow another operator's bandwidth during congestion periods) would be difficult to realize. Therefore, static resource sharing remains the only viable solution.

3. The European Telecommunications Standards Institute (ETSI) Telecommunications and Internet-converged Services and Protocols for Advanced Networks (TISPAN) initiative is a Next Generation Network (NGN) architecture focusing on fixed networks and Internet convergence.

The 3GPP PCC architecture has limited FMC support in the home-routed roaming scenario (Figure 4.3b). Because the P-GW is located in the home network, handovers requiring BBERF relocation can suffer excessive delays and as a result lead to a bad user experience. Additionally, the visited network does not have a significant role to play in the selection of the AEG, although the visited operator has an up to date knowledge of its network's condition.

The RACS (Figure 4.4) is handicapped regarding global mobility, FMC and roaming because the user is tied to the Access Network (AN) by the Network Attachment Sub-System (NASS). This also complicates the integration of Mobile Virtual Network Operators (MVNOs).[4] However, the RACS isolates service policies from network management (which can distinguish admission control in the access and aggregation segments).

On the bright side, both policy architectures clearly separate the services from the network.

5.4.2 Requirements on PCC to Better Support CN Sharing and FMC

In [46] are listed four requirements that must be met in order to maximize an operator's flexibility to follow any desired path (e.g., broadening its palette of accesses or growing into advanced infrastructure sharing scenarios). FMC and CN sharing have no link *a priori*, but an evolution of the PCC architecture improves the support of both aspects if it meets the requirements below:

1. services must be completely decoupled from the network. This implies that network policies and resource management, on one side, are *independent* from user and/or service policies on the other side;
2. every business/network entity must control its resources following a client/server model;
3. network management must be decoupled from subscriber management (i.e., service and user policies, IP address allocation, authentication). This eases the integration of a MVNO on top of an existing CN;
4. for better FMC roaming support, the AEG should be selected by the visited network because the home network knows nothing about the visited network's condition.

The evolved PCC architecture must also be capable to support combined gateway scenarios (in which the P-GW and S-GW are merged into a single node) as well as the presence of several P-GWs (resulting from CN sharing and/or large operators having multiple peering points).

4. A MVNO has no network infrastructures nor spectrum license but it has a subscriber base and can offer services to its subscribers.

5.4.3 Overview of the Solution

As a first step to present our solution (Figure 5.1), we define business roles that separate the responsibilities between the business/network entities [5]:

- the **Network Service Provider (NSP)** hosts AFs (IMS [6] or other), authenticates and bills users, gives access to IP networks beyond the EPS (e.g., the Internet). The NSP dictates service- and user-related policies (hosts a PCRF and a SPR), controls the border gateway functions of the P-GW and interacts with remote NSPs for roaming. All NSPs are MVNOs in our study;
- the **IP Aggregation Network (IPAN)** provider offers IP transport and infrastructure services (e.g., antivirus DPI [7] engine) to the NSP(s) according to any Service Level Agreement (SLA) in force. The IPAN links NSPs to ANPs, also hosts EPC edge nodes (e.g., S-GW, P-GW) and hides the details of the underneath transport architecture;
- the **Access Network Provider (ANP)** manages the resources of an access network according to a set of local policies and enforces the SLAs made with the IPAN operator(s). Note that our model does not forbid ANPs from interacting with several IPANs.

As shown in Figure 5.1a, the roaming architecture always uses the visited network's P-GW as a mobility anchor. For the home-routed scenario, this reduces the signaling delays during micro-mobility involving a BBERF relocation. It also enables the visited network to actively participate to the AEG reselection process (e.g., due to changing network conditions). Finally, this scenario could allow the home network to benefit from the IPAN's infrastructure services.

The concept of infrastructure services encompasses all of the advanced functions (e.g., DPI antivirus, audio/video transcoding, file caching, legal interception) that an IPAN offers to NSPs. They provide value-added services for both IMS and non-IMS applications. [8] An evolution of the S9 (inter-operator) reference point could provide all of the required infrastructure services discovery and invocation mechanisms.

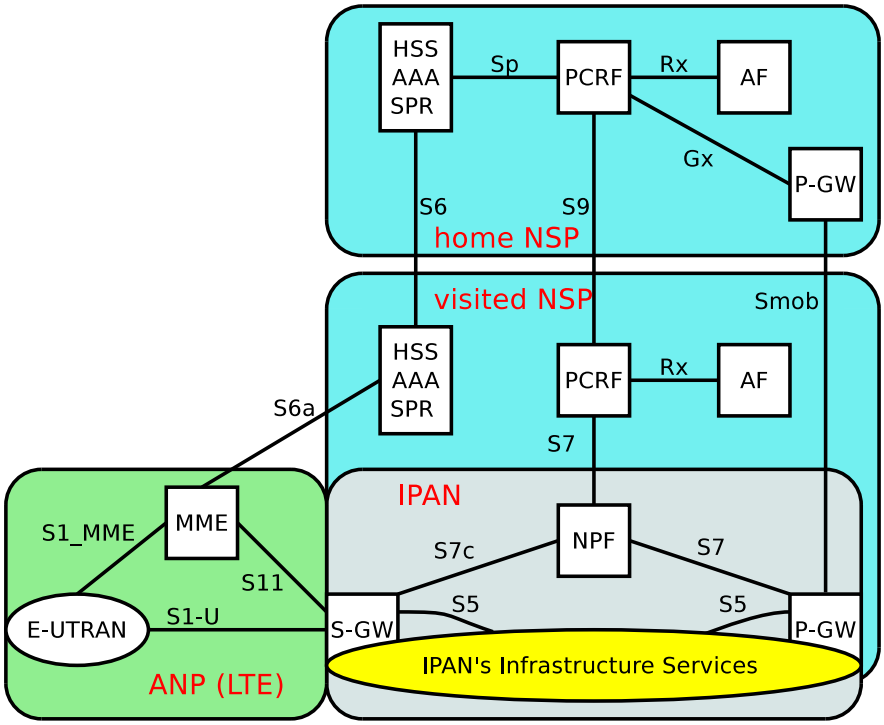
The evolved architecture (Figure 5.1) was slightly refined since [46] in order to reconciliate both PCC architecture variants (based on either GTP or PMIPv6) and minimize the impacts on the P-GW. As a matter of fact, a reference point linking a PCRF to its P-GW was removed; it was decided that the NPF would transparently forward the charging information to the P-GW.

5. Note that traditional operators play all roles.

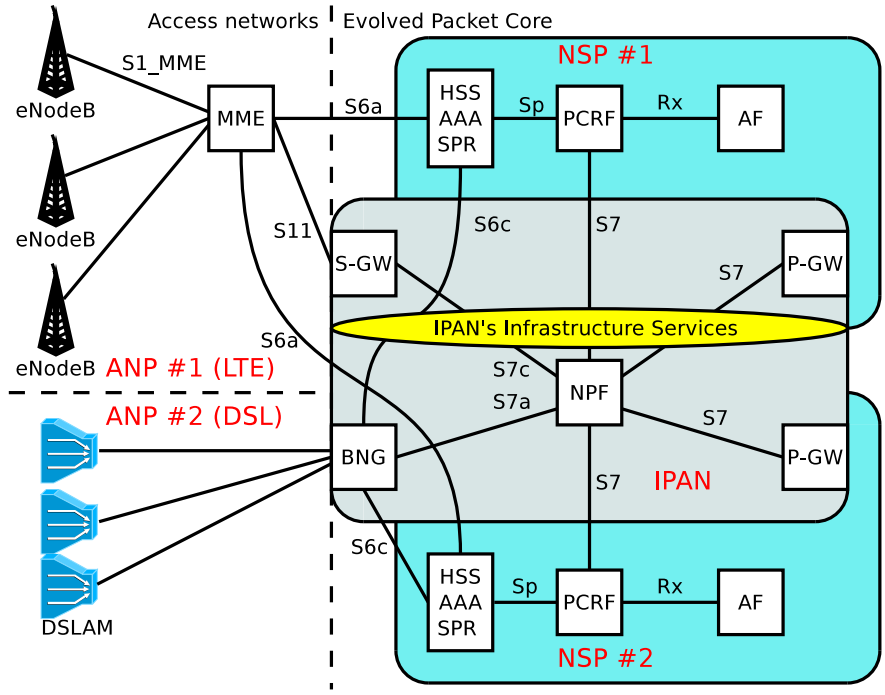
6. The IP Multimedia Subsystem (IMS) is an architectural framework for delivering IP multimedia services.

7. Deep Packet Inspection (DPI) implies that the protocol headers as well as the data traffic are inspected.

8. The location of these infrastructure services is intentionally vague as some, notably most location-based services, should be implemented closer to the UE.



(a) Roaming architecture valid for both *home-routed* and *local breakout* cases (control and data planes).



(b) *Non-roaming* architecture featuring CN sharing and multiple accesses (control plane only).

Figure 5.1 Studied evolution of the 3GPP PCC and QoS architecture.

The evolution of the 3GPP PCC architecture we study combines the key features from both the 3GPP PCC and TISPAN RACS policy engines while avoiding their respective drawbacks. Our study reveals that a 3GPP/TISPAN hybrid solution can greatly improve the support of FMC and CN sharing while minimizing the impacts on the other CN nodes. Consequently, we introduced the Network Policy Function (NPF) between the PCRF and the P-GW (see Figure 5.2).

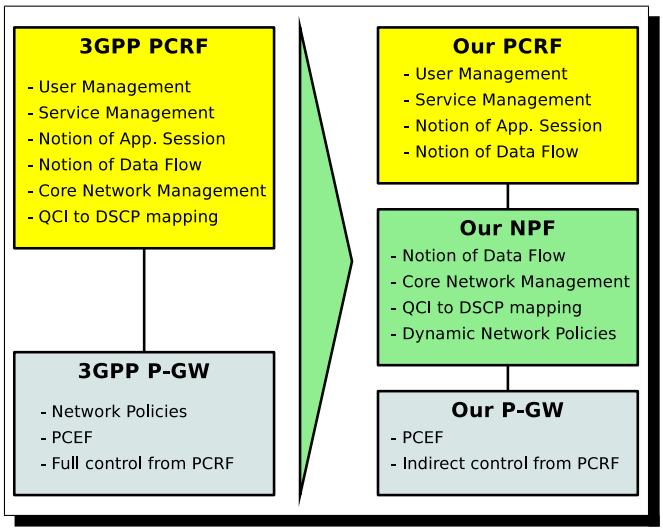


Figure 5.2 Functionality assignments to the CN nodes, comparing 3GPP and our study.

The NPF determines network policies [9] and shares the IPAN resources (including infrastructure services) among NSPs. It performs a coarse bandwidth management as opposed to the finer (i.e., per SDF) resource management done by the P-GWs. The NPF hides the details of the underneath transport network (e.g., MPLS-TP, PBB-TE) from NSPs. As a result, the IPAN appears as a plain DiffServ IP network to NSPs.

If we suppose for a moment that infrastructure services are not implemented into the IPAN, then our solution is readily compatible with both PCC architecture variants. Indeed, the existing Gx reference point found between the PCRF and P-GW (see Figure 4.1) would be reused in a chain to link the PCRF, the NPF and P-GW. Furthermore, for the PMIPv6-based IPAN, the BBERF located in the AEG would receive its QoS rules directly from the NPF over the Gxx reference point; the NPF extracts the information needed by the AEG from the PCRF's request.

The number of devices that can actively participate to a Generalized MPLS (GMPLS) control plane is expected to increase in the future. As a result, the NPF constitutes a key lo-

9. An example of network policy could be that a minimum of 25% of the network resources must be reserved for best-effort traffic.

cation to implement centralized network management functions such as a Traffic Engineering Database (TED) and a master Path Computation Element (PCE). Indeed, Traffic Engineering (TE) allows an IPAN operator to balance the traffic load among the network elements and ensure that each traffic flow aggregate meets its QoS requirements at all times.

Using a single IP infrastructure for all communication needs (telephony, emergency services, internet/corporate network access, text messaging, entertainment, etc) imposes stricter requirements on the IPAN's tolerance to link or node failures. As an example, MPLS-based IPANs could benefit from PCE-assisted TE to compute backup paths that are not part of the same Shared Risk Link Group (SRLG) than their primary Label-Switched Paths (LSPs).

5.5 Changes to the PCC Nodes

As a consequence to introduce the NPF in the PCC architecture, some functions of the PCRF and P-GW were relocated into the NPF. These changes were required in order to adequately separate subscriber management from CN management. This section provides an overview of the impacts our solution has on the CN nodes.

5.5.1 Impacts on the P-GW

The P-GW sits on the border that separates the IPAN from a PDN.[10] Each NSP hosts at least one P-GW and many NSPs can share an IPAN.

As shown in Figure 5.3, the NPF adds to PCC rules the information a P-GW needs to uniquely identify a physical or logical link toward the AEG. Additionally, the NPF can specify L2/L3 marks to apply to the all forwarded packets. These marks can specify L2/L3 QoS classes, dropping precedences or can be used to separate traffic from distinct NSPs.

When infrastructure services are not considered, the only modification needed to the Gx reference point is the removal of the `ToS-Traffic-Class` Attribute-Value Pair (AVP).[11] Indeed, the NPF needs only the QCI to identify the QoS class. On the other hand, the introduction of infrastructure services needs to define a number of AVPs for service requests. It is out of the scope of this paper to determine if the infrastructure services AVPs should be standardized or operator-defined.

In our study, the PCEF is an IPAN-controlled entity. However, some other functions of the P-GW are definitely to be controlled and configured by the corresponding NSP. Following are some examples of these functions:

– IP address allocation;

10. A P-GW can give access to many PDNs; each PDN is uniquely identified by its Access Point Name (APN).

11. Like most of the reference points defined in the EPS, Gx is based on the Diameter protocol.

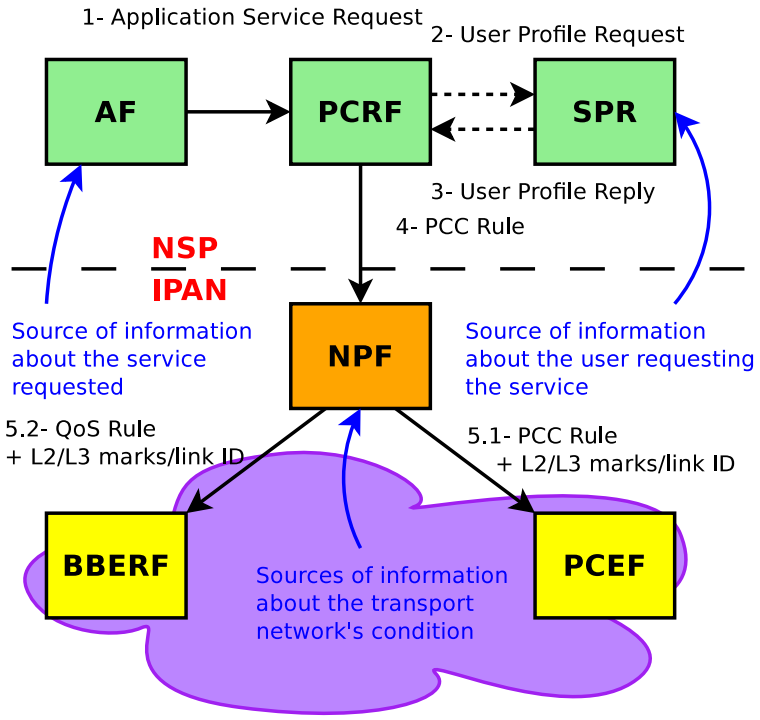


Figure 5.3 Policy decision process for the IETF-based EPC showing the information sources. A GTP-based EPC would feature no BBERF.

- interfacing the OFCS and OCS;
- accounting for inter-operator charging (e.g., when a roaming UE uses a visited NSP's application);
- gating control (packet filtering);
- Dual Stack Mobile IPv6 (DSMIPv6) Home Agent if the S2c reference point is used.

Some functions like Lawful Interception can be implemented in a few different places but a country's regulations may set requirements on the location or legal responsibility of the services.

5.5.2 Impacts on the S-GW

In the evolved roaming architecture (Figure 5.1a), the S-GW tunnels the UE's traffic to the visited P-GW in all cases. As a result, the S-GW needs not support any traffic accounting because this task is now always performed by the P-GW.

PMIPv6-based EPC

In our solution's PMIPv6 architecture variant, the EPC edge nodes [12] control all user traffic that traverses the IPAN. Consequently, PMIPv6 mobility management is fully under IPAN control.

Additionally, our solution eliminates the need for the S-GW to support S8-S2a/b chaining [13] because all traffic to/from non-3GPP accesses traverses the visited network's P-GW, as in the non-roaming scenario. As a result, the PMIPv6 Local Mobility Anchor (LMA) function can be completely removed from the S-GW.

Furthermore, if the EPC edge nodes support a tunneling mechanism (e.g., MPLS) that can carry packets having the same QoS needs, then it becomes possible to implement PMIPv6 header compression between the EPC edge nodes. Indeed, PMIPv6 user plane tunneling (see Figure 5.4) first encapsulates a user's IP packet into Generic Routing Encapsulation (GRE), with the GRE key option used to uniquely identify the PDN connection, then into either an IPv6 or IPv4 packet (with optional UDP encapsulation). [14] In this particular case, a LSP's endpoints specify the LMA and Mobility Access Gateway (MAG) addresses, the tunnel's QoS class is specified by DiffServ Code Point (DSCP). As a result, the edge nodes could strip the outer IP header because its contents has now become redundant.

Finally, the S8 reference point toward the home P-GW is replaced by the Smob reference point (between the home and visited P-GWs). Smob plays the same role as S8 but reuses the P-GW as the top level mobility anchor.

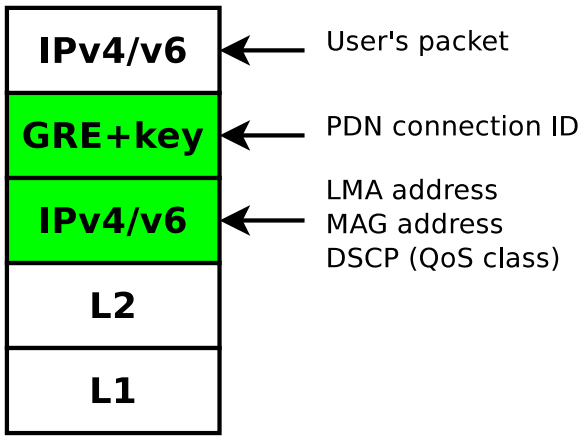


Figure 5.4 PMIPv6 User Plane Stack.

12. Collectively referring to P-GWs and AEGs.
13. S8-S2a/b chaining refers to a capability of the S-GW in the visited network to act as a local mobility anchor between 3GPP and (un)trusted non-3GPP accesses [6].
14. User Datagram Protocol (UDP) encapsulation is used for Network Address Translation (NAT) traversal.

5.5.3 Impacts on the PCRF

In the 3GPP PCC architecture a PCRF takes decisions based on three categories of operator-defined policies considering:

1. the service being requested (from the AF, over the Rx reference point);
2. the profile of the subscriber (from the SPR via the Sp reference point);
3. the conditions of the transport network (inputs fed to the PCRF from various sources).

In our solution, a PCRF bases its policy decisions on the same inputs as a 3GPP PCRF (see Section 5.2). However, network load is evaluated according to the NSP's limited view of the physical infrastructures, as presented by the NPF.

Additionally, our PCRF isn't affected by the choice of architecture variant (GTP/PMIPv6) implemented into the IPAN. Indeed, only a S7 reference point links the PCRF to the NPF located into the IPAN (refer to Figure 5.1).

Finally, even though the S7 reference point is based on the existing Gx interface (which links the PCRF and the P-GW in the 3GPP PCC architecture), S7 needs to be augmented with many yet to be defined AVPs in order to allow the invocation of IPAN infrastructure services.

5.5.4 Impacts on the MME

The Mobility Management Entity (MME) is a control plane node that plays a central role in user authentication, authorization, mobility and session management for the 3GPP RANs (GERAN, UTRAN and E-UTRAN) as well as with HRPD.[15] The MME is also responsible for tracking and paging idle UEs, manages all bearer procedures and selects the gateways (S-GW and P-GW) in a UE's initial attachment procedure.

Our study considers the MME as an access-specific node because it is needed exclusively for HRPD and 3GPP accesses. Furthermore, the MME constitutes a privileged location for implementing an access network's policy server.

In both roaming scenarios, our solution imposes that the MME communicates with the visited network's Home Subscriber Server (HSS) which acts as a relay to the home HSS. Moreover, our solution's home-tunneling scenario uses the visited P-GW as a mobility anchor. As a consequence, the home HSS needs not be aware of the UE's micro-mobility events, thus reducing the number of signaling messages between the home and visited networks.

15. High Rate Packet Data (HRPD) is a RAN technology similar to 3GPP's UTRAN that was developed by the 3rd Generation Generation Partnership Project 2 (3GPP2).

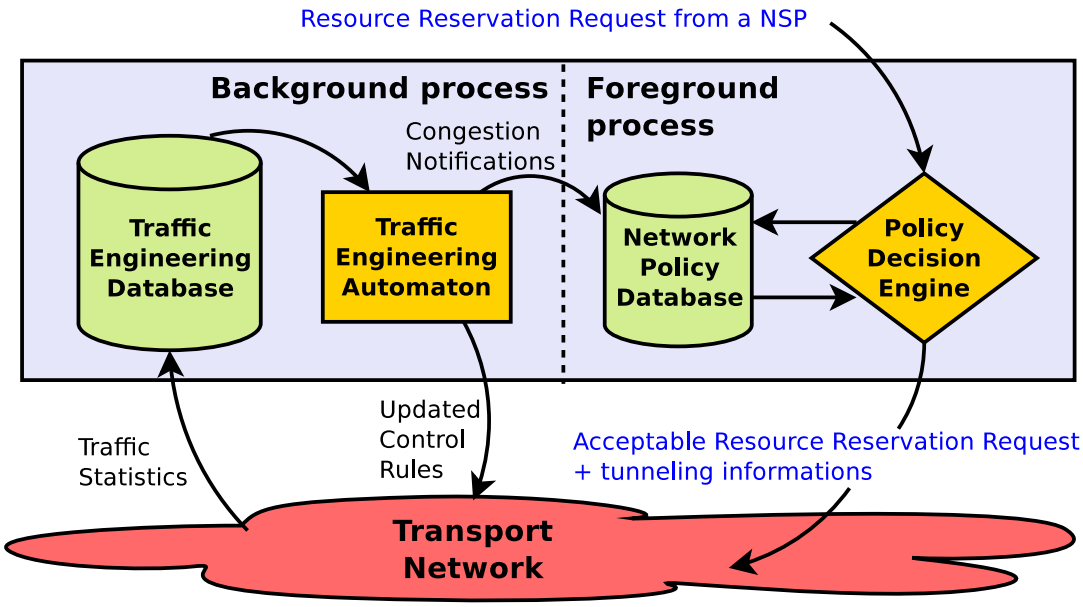


Figure 5.5 Network Policy Function (NPF) internals with optional traffic engineering support.

5.6 Details of the NPF

The NPF is a policy decision point whose main goal is to separate the management of the IPAN from both user and service management that is done by the NSPs. As a consequence to introduce the NPF in the CN, all NSPs are MVNOs in the evolved PCC architecture we study.

Figure 5.5 details the internals of a NPF node that performs traffic engineering over some L2 tunneling mechanism used in the IPAN. Only the foreground process is visible to a NSP; the background process (if implemented) collects traffic statistics from the IPAN's network elements. Updated control rules can be used to steer traffic flows around a failed link or for load balancing. These control rules can also affect the network policies database, allowing them to dynamically adapt to the IPAN's ever changing conditions.

Some network policies arise as a consequence of the IPAN's topology and of the selection of a tunneling mechanism in particular. For example, a network policy could be that the sum of the bandwidths allocated to some L3 QoS classes must not exceed a given total because they are mapped to a single L2 QoS class below.

CN sharing scenarios force the NPF to ensure that the SLAs contracted with each hosted NSP are respected at all times. Furthermore, the NPF is useful for hiding the IPAN's complexity from the NSP. This is true even for a non-shared CN.

Following a resource reservation request from a PCRF, the NPF first ensures that the request conforms to the SLA in force between the NSP and the IPAN operator. Then, the

NPF identifies a LSP that offers adequate QoS and for which no congestion notification was received. The ARP of the request can influence the path selection. If the request is accepted, the NPF appends the tunneling information and/or L2/L3 mappings before forwarding it to the PCEF.

The solution we studied follows a principle inspired from TISPAN's RACS; network resources in the IPAN and each AN are managed independently. This implies that a resource reservation received from a NSP is only accepted if both the IPAN and the AN have sufficient resources. [16] Additionally, the NPF and AEGs interact using any required access-specific mechanism in order to provide a SDF with the appropriate QoS.

The concept of *infrastructure services* was introduced in the IPAN in order to enable some bandwidth optimizations that benefit the NSPs or to improve the users' QoE. These services allow the IPAN to become much more than just a dumb bit pipe. The NPF plays a key role for controlling how NSPs share these network resources. A detailed example of a Video On-Demand application featuring a caching server collocated with the S-GW was presented in [46].

5.7 Conclusion

This paper presented the details of a potential evolution of the 3GPP PCC architecture that is compatible with both of its architecture variants (based on GTP or PMIPv6). We showed how our solution allows NSPs to evolve toward either more flexible FMC support and/or more elaborated CN sharing scenarios by respecting the requirements listed in Section 5.4.2. This allows operators to meet the CN's future needs.

We have also demonstrated that the studied PCC roaming architecture would be significantly simplified when compared to the existing 3GPP PCC architecture. Indeed, our solution eliminates a number of redundant mechanisms from the CN nodes and reuses as much as possible the features that are already present other CN nodes. We thoroughly described the impacts our works have on every CN node.

Finally, we defined the concept of infrastructure services as value-added services that turn the IPAN into a smart bit pipe. Further study is required to extensively describe how infrastructure services will be announced and invoked.

16. Our model allows ANPs to do business with more than one IPAN and to dictate their own access policies.

CHAPITRE 6

A MULTIACCESS RESOURCE RESERVATION PROTOCOL FOR THE 3GPP EVOLVED PACKET SYSTEM

Auteurs : Stéphane Ouellette et Samuel Pierre.
Revue : Soumis au *Journal of Computer Science (Science Publications)* le 11 avril 2012.

Abstract

Starting with Release 8 onward, the 3rd Generation Partnership Project (3GPP) standards specify the Evolved Packet Core (EPC) which is an IP-based Core Network (CN) for cellular networks, capable of supporting high-speed real-time packet services. The Evolved Packet System (EPS) is the union of the EPC and the Evolved Universal Terrestrial Radio Access Network (E-UTRAN), the latter features the Long-Term Evolution (LTE) pre-4G radio technology.

In addition to the widely deployed GSM, W-CDMA, LTE and CDMA2000 accesses, the EPC supports accesses such as Wi-Fi and WiMAX; others could be added later. However, the association of the CN's Quality of Service (QoS) classes to those of the accesses are not trivial unless the assignments are static. Moreover, with the convergence of services as a business trend, it is important that the QoS offered to the applications be managed in a uniform way.

On the other hand, the current specifications make it difficult for new applications that are not supported by the network to request some QoS support by the network infrastructures. The problems mentionned above motivate the creation of an access independent, future-proof multiaccess resource reservation protocol.

6.1 Introduction

After three decades of existence and expansion, the mobile communications industry has recently entered a stabilization phase that characterizes a mature industry. Operators now focus on cost reductions because their profit margins shrink. Among the current business trends (see [46] for more details) is the convergence of networks, services and terminals, globally refered to as Fixed-Mobile Convergence (FMC).

FMC aims to offer IP networking and telephony on a single device that can switch between local- and wide-area networks. It enables network operators to offer their users either the

best or the most cost effective available connection on all accesses. FMC helps operators to reduce their costs as a single CN is shared between all accesses and is an added value to the customers because it allows them to save of connection costs and benefit from an usually faster local access point.

However, future applications will require a QoS to function adequately. Within 3GPP networks resources are managed by the Policy and Charging Control (PCC) architecture [2]. Services that need QoS host an application server located in the operator's network. This server interacts with the PCC architecture's policy server in order to request network resources.

On the other hand, new or unsupported applications face the chicken or the egg problem: they will not function satisfactorily (and hence will not be widely deployed) because they cannot request QoS from the network and they will not receive QoS unless they are supported by it. This problem is aggravated by the number of different accesses they must support, because each access defines its own QoS mechanisms.

This paper is organized as follows. Section 6.2 presents an overview of the PCC architecture and explains how QoS is supported in the EPC. Section 6.3 decribes in details the QoS issues for unsupported applications in 3GPP multiaccess networks and lists some requirements that a potential solution must satisfy. Section 6.4 describes our proposal and includes a diagram that illustrates the sequence of operations. Section 6.5 discusses the advantages of our solution. Section 6.6 presents the model of a complete system that we formally verified with the help of the Uppaal model checker. Finally, Section 6.7 summarizes our contribution and describes our future works.

6.2 3GPP PCC Architecture and QoS Management in the EPS

The EPS [5, 6] is basically composed of a radio access network (the E-UTRAN) and of the EPC which is a multiaccess CN. Our intention is to present an overview of the multiaccess EPC variant [6] and provide the necessary background for understanting the context of our work. Refer to [15, 47] for details on the PCC architecture and its QoS management procedures.

As illustrated in Figure 4.1, the E-UTRAN is made of a single kind of node: the LTE base station (eNodeB). The eNodeB manages the wireless interface between itself and the terminals, including QoS for all terminals and their active sessions in both the uplink and downlink directions.

On the EPC's control plane, the Policy and Charging Rules Function (PCRF) is the logical node that takes policy decisions based on the network conditions, the service requested and

the user's profile. The Application Function (AF) (hidden in the "operator's services" cloud) is a server that runs an application and requests network resources from the PCRF on behalf of the user. The user's profile is retrieved from the Subscription Profile Repository (SPR) and contains a list of applications and QoS informations applicable to the subscriber. Finally, the Mobility Management Entity (MME) handles all control plane signaling, including mobility and security functions for the terminals.

On the data plane, the Policy and Charging Enforcement Function (PCEF) performs QoS enforcement, gating control, online and offline charging. It is implemented into the Packet Data Network (PDN) Gateway (P-GW) that separates the operator's network from the EPC. Finally, the Bearer Binding and Event Reporting Function (BBERF) is implemented into the Access Edge Gateway (AEG) for each access type. [1] Its main tasks are event reporting to the PCRF and *bearer binding*. [2]

Compared with the previous CN generations, QoS setup has been greatly simplified in the EPS. The EPS introduced the concept of QoS Class Indicators (QCIs) to identify QoS profiles. Each QCI specifies a set of QoS parameters (maximum delay and jitter, scheduling weight and maximum packet error loss rate). An EPS must support at least nine standardized QCIs.

6.3 Problem Statement

Several issues need to be addressed in the EPS to facilitate QoS management in a context where FMC, mobility and dynamic network policies are simultaneously considered. Following a description of these issues, we list the requirements that a solution must satisfy, taking in consideration all of the design constraints.

First, 3GPP accesses prior to LTE support QoS with a terminal-initiated procedure known as "**secondary Packet Data Protocol (PDP) context activation**" that is used to setup a radio bearer with the specified QoS properties. The number of combinations of QoS parameters is large, making this mechanism hard to implement in the network nodes. Additionally, as it is difficult for application developers to determine (or "guess") a dozen of QoS parameters, it often leads to an erroneous QoS setup.

LTE supports an equivalent of the "**secondary PDP context activation**" mechanism. However, some terminals may not support this mechanism and operators exhibit a strong preference toward network-initiated QoS because this gives them full control over the network.

Second, it is easy for an application server hosted by the operator's network to setup QoS

1. For the multiaccess variant of the EPS, the BBERF is part of the Serving Gateway (S-GW).
2. A bearer is a logical IP data path between the terminal and the network with specific QoS properties. Bearer binding is the association between a Service Data Flow (SDF) and the network bearer.

by sending a resource reservation request to the PCRF. However, applications that are not supported must rely on terminal-initiated QoS (which may not be allowed or supported) or function without QoS guarantees.

Third, *vertical handovers*[3] introduce the problem of correctly translating the QoS parameters between different access technologies. Moreover, QoS support can be implemented as resource reservations or service differentiation (or a combination of both paradigms, depending on the type of access).

Fourth, there must be a mapping of the 3GPP QCIs to the underneath IP transport network's QoS classes as well as to each access technology. Because the 3GPP architecture's only design requirement is to to sit on top of an IP transport network, the association of 3GPP QoS classes to the IP level QoS classes is trivial if these assignments are statically configured in both the network elements and the terminal. This can be a problem for roaming users if the mappings differ from an operator to another. At last, dynamic mappings can be based on other parameters that are not considered at the moment. [4]

Finally, the existing Internet Engineering Task Force (IETF) resource reservation protocols[5] operate end-to-end and use a hop-by-hop mechanism that forces all supporting routers to process the protocol messages. These protocols use the IP Router Alert Option [32, 33] which is considered by many as a potential security hazard because it is often processed on a router's *slow path*[6] Additionally, QoS in the EPS does not extend beyond the P-GW and needs to be authorized by the PCRF. Moreover, some of the existing protocols' features are useless in our case, e.g., NSIS' network discovery mechanism. At last, these protocols are not yet compatible with 3GPP's QCIs.

6.3.1 Requirements

As discussed earlier, the different types of accesses to be supported by the EPC define their own QoS mechanisms. Also, some accesses need several QoS parameters to be specified in order to adequately setup QoS. Thus, as a first requirement, an application must define QoS solely in terms of QCIs and rely on the network to specify the corresponding QoS parameters.

We also mentionned that operators favor network-initiated QoS and that applications not supported by the network must rely on terminal-initiated QoS or execute without QoS. Therefore, the second requirement is that the desired solution must be described as a "ter-

3. A *vertical handover* implies a change of access.

4. As an example, two users on a Wi-Fi access could get a different *Transmit Opportunity* (time allowed for transmitting) depending on their identity.

5. Examples are Resource reSerVation Protocol (RSVP) [13] and Next Steps In Signaling (NSIS) [23].

6. The *slow path* refers to a general Central Processing Unit (CPU) that processes all exceptional IP packets, by opposition to the *fast path* which processes all ordinary packets with fast dedicated hardware.

minal-triggered network-controlled QoS" mechanism.

Vertical handovers support require that the network translates the application's needs (expressed in terms of 3GPP QCIs) into QoS parameters suitable for the new access point everytime a vertical handover is performed. Additionally, the mappings from 3GPP QCIs to the underneath IP transport network's QoS classes enable the terminal to classify and/or mark the packets waiting in its transmission queue.

The desired solution shall be used only in an EPS. Every resource reservation request must be processed by the PCRF. As a result, the protocol messages must be exchanged exclusively between the terminal and the Bearer Binding Function (BBF), which will be implemented into either the P-GW (for the variant described in [5]) or the S-GW (for the multiaccess EPC variant [6]). [7]

Finally, the protocol messages will not be routed across a network: it must be a point-to-point protocol. This requirement also implies that the desired solution must not use the IP Router Alert Option [32, 33] and does not implement a network discovery mechanism.

6.4 Description of MARSVP

The proposed MultiAccess Resource ReSerVation Protocol (MARSVP) is access-independent by design but enables the network to set access-specific QoS configuration parameters. The protocol lets the network operator figure out how to accommodate the resource reservation beyond its own network. QoS as such is guaranteed on the access segment but can be fully or statistically guaranteed beyond the access segment.

The protocol messages are described below:

- `RESERVE` (*terminal to access router*): Contains a unique session identifier and, for each data flow described, the 5-tuple packet filter (addresses/ports/protocol), guaranteed and maximum bit rates, 3GPP QoS QCI and a priority level (relative importance of a flow).
- `ACCEPT` (*access router to terminal*): Informs the terminal that a reservation was (fully or partially) accepted. Flow descriptors contain the accepted bandwidths for each accepted flow. L3/L2 QoS classes mappings are included for both uplink/downlink flows. If the allowed bandwidth is zero, then a given flow's reservation request was denied.
- `ERROR` (*access router to terminal*): Informs the terminal that an error (protocol, syntax, authorization, etc) has occurred and that the resource reservation has failed. If sent in response to a `MODIFY` message, the existing resource reservation is not affected by the

7. The BBF is the first node on the data path that interacts with the PCRF.

reception of the ERROR message.

- REFRESH (*terminal to access router*): On accesses that do not provide terminal detection, reservation requests must be periodically refreshed. Contains only the session identifier.
- MODIFY (*either ways*): Contains only the flow descriptors that must be changed. A bandwidth of zero means that a given flow's reservation is torn down.
- TEARDOWN (*either ways*): The whole session must be torn down. Contains the session identifier and the reason of the teardown.
- TEARDOWN_ACK (*either ways*): Acknowledges the reception of a TEARDOWN message.

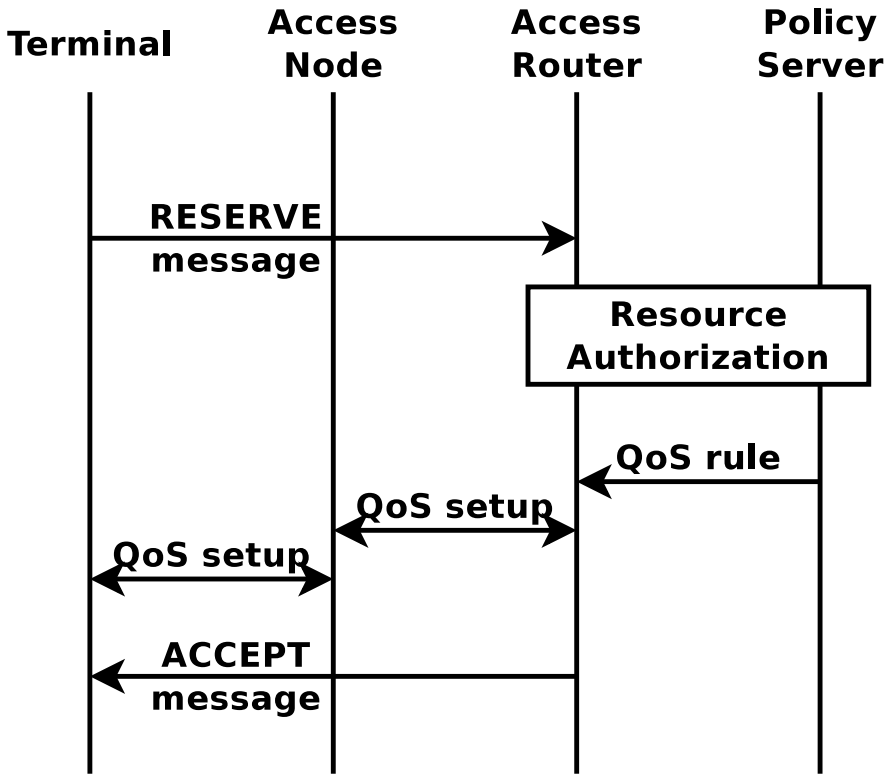


Figure 6.1 Resource reservation with MARSVP

A successful initial resource reservation is illustrated in Figure 6.1 and described below.

Step 1: A fixed or mobile terminal sends a RESERVE message to the access router or first IP node to reserve some resources based on a session identifier and, for each data flow described, a packet filter (addresses/ports), guaranteed and maximum bit rates, 3GPP QCI and a priority level useful in congestion situations. [8]

Step 2: The access router proceeds with a regular set of actions between itself and the PCRF as per existing 3GPP PCC procedures.

8. The 3GPP Allocation and Retention Priority (ARP) specifies the relative importance of a data flow.

Step 3: For each accepted flow, the PCRF can send a QoS rule to the access router. This allows QoS to be set up for the terminal. When QoS is enabled, an `ACCEPT` message is sent to the terminal by the access router. However, if the request is denied (out of resources, request rejected by the policy server, etc), an `ERROR` message is sent to the terminal.

Step 4: Following an `ACCEPT` message, the data transfer takes place. Periodic `REFRESH` messages may be sent by the terminal if specified in the `ACCEPT` message.

Step 5: When the terminal is done with the data transfer, it sends a `TEARDOWN` message to request the termination of the resource reservation. The access router acknowledges with a `TEARDOWN_ACK` message. Then, the access router terminates the resource reservation.

A `MODIFY` message can be sent by the terminal in the data transfer period to request a QoS modification for the specified data flow(s). An `ACCEPT` or `ERROR` message is sent by the access router in response to a `MODIFY` request. If an `ERROR` message is sent, the existing resource reservation remains valid. If more than one data flow modifications were requested, an `ACCEPT` message can partially accept the modifications (rejected modification are indicated by a bandwidth of 0, the existing resource reservations are unaffected).

A `MODIFY` message sent to the terminal by the access router informs the terminal that the conditions of the network have changed and QoS is updated. The terminal doesn't acknowledge the `MODIFY` message.

A `TEARDOWN` message sent to the terminal by the access router immediately terminates the whole session (and all of the data flows). A reason code indicates what caused the session to be torn down.

6.5 Discussion

There are many advantages to MARSVP. First, a 100% software solution is easier to implement on existing equipments. A generic QoS Application Programming Interface (API) will be defined for the terminal that allows simplified (and unified) QoS management. MARSVP uses UDP/IP over the terminal's data channel and is thus portable to any type of access.

Second, if the current access can detect the presence of the terminal, MARSVP can optionally be "hard state" (no need to refresh the resource reservations at periodic intervals to save link bandwidth). Otherwise, reservations need to be periodically refreshed by the terminal with the help of the `REFRESH` message.

Finally, the QoS needs of an application are specified using an uniform syntax. The network is responsible to translate these specifications into adequate QoS parameters for the current access. Should the terminal perform a vertical handover, the network would resend

an `ACCEPT` message with the adequate QoS mappings.

6.6 Formal validation of the state machines of MARSVP

Following the functional description of MARSVP and considering the lossy nature of wireless communications, it is important to ensure that the state machines of all of the entities involved in MARSVP's operations be able to cope with the loss of any message.

We used the Uppaal v4.0.13 model checker to validate our protocol. States are represented with small circles called *locations*. Blank locations indicate states in which it is possible to wait for an external event (called a *signal*) or a timeout. An *invariant* is a condition associated to the state that must remain true in order to be allowed to wait in that state. If the invariant becomes false, the state machine must immediately proceed with a state transition.

Locations marked with a "U" are said to be *urgent*; it is not possible to wait in that state and a state transition must be made without delay. However, other state machines of the model are allowed to work. Locations marked with a "C" are said to be *committed*. Committed states behave like urgent states but forbid any transitions in the other state machines of the model.

State transitions are specified with arrows. A *guard* is a condition to enable a state transition. As a consequence to a state transition, it is possible to update one or more variables of the model (e.g., a counter, timer, etc).

Signals allow state machines to communicate. They are sent or received during state transitions. An exclamation mark appended to a signal name indicates that the given signal is emitted during the state transition. On the other hand, a question mark appended to a signal name specifies that the associated state transition can only be made when the signal is received.

In order to take into account the lossy nature of wireless communications, we duplicate the state transitions that send a protocol message to another entity over a wireless link. That way, we can demonstrate that our state machines can survive the loss of a message and avoid losing their synchronization. As depicted in Figure 6.2, the upper state transition simulates the loss of the message while the lower state transition represents a successful transmission.

Our model includes 3GPP's BBF as well as the User Equipment (UE). The UE is logically separated into the `Terminal` device itself and the `Application` that runs over it. Figure 6.3 illustrates the `Application`'s state machine. The `Terminal` device and the `AccessRouter` are respectively depicted in Figure 6.4 and Figure 6.5.

Figure 6.2 Example showing how to simulate the loss of a message

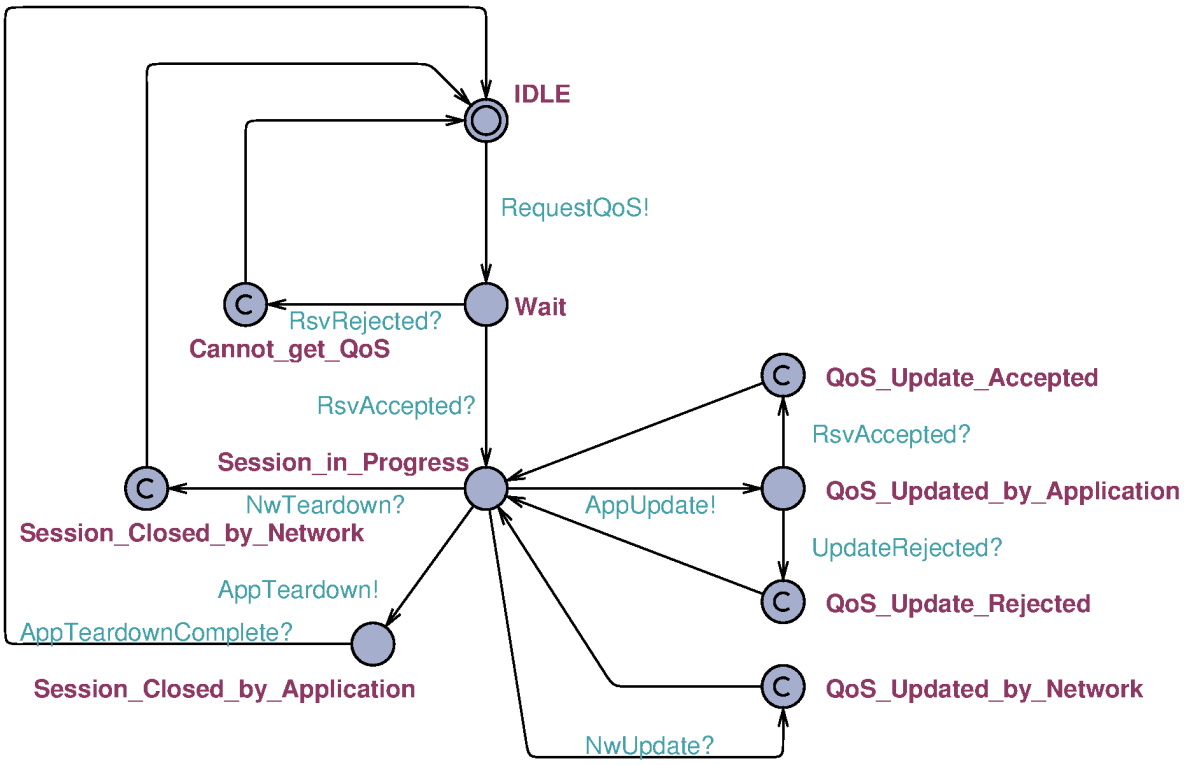


Figure 6.3 State Machine of the Application

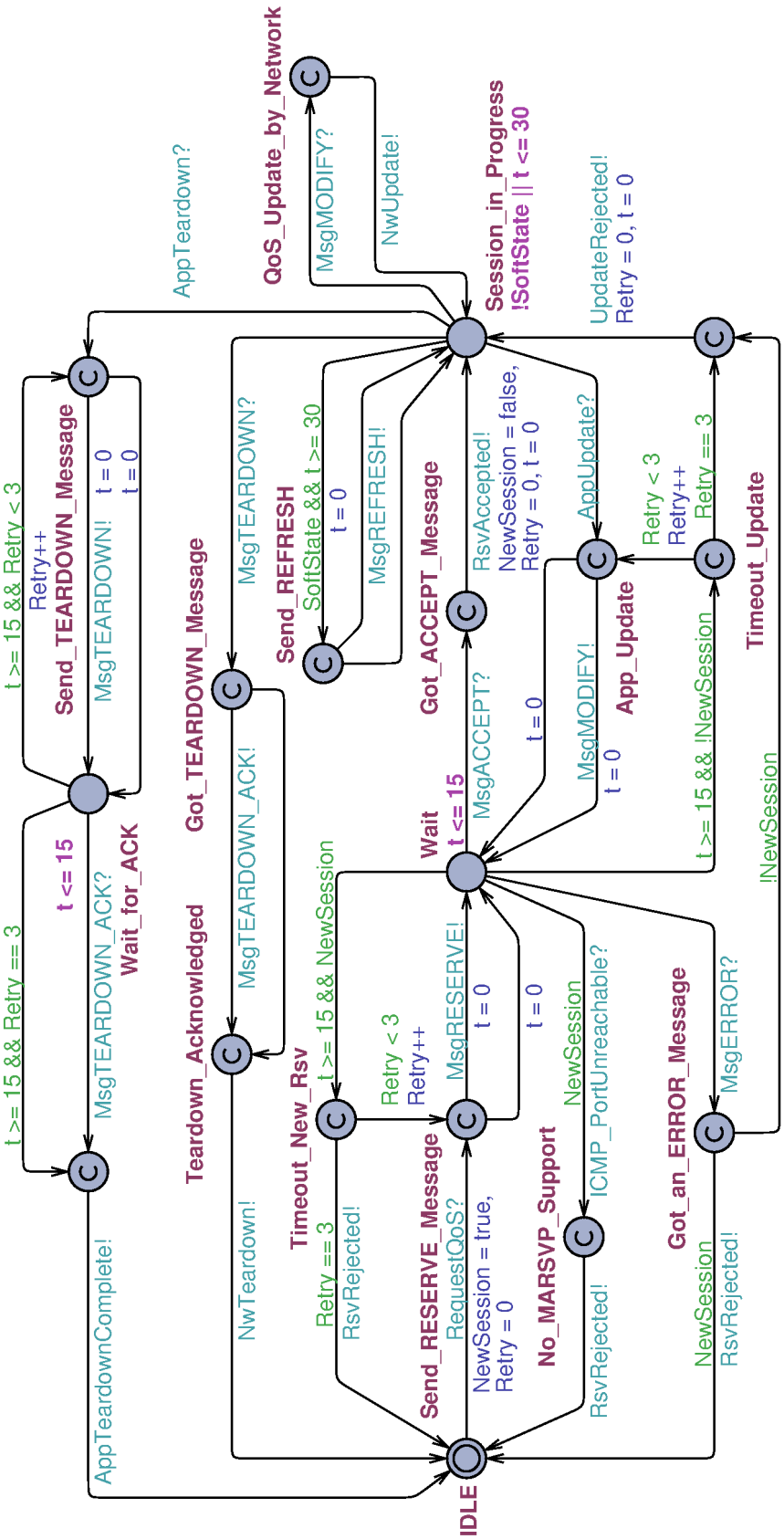


Figure 6.4 State Machine of the Terminal

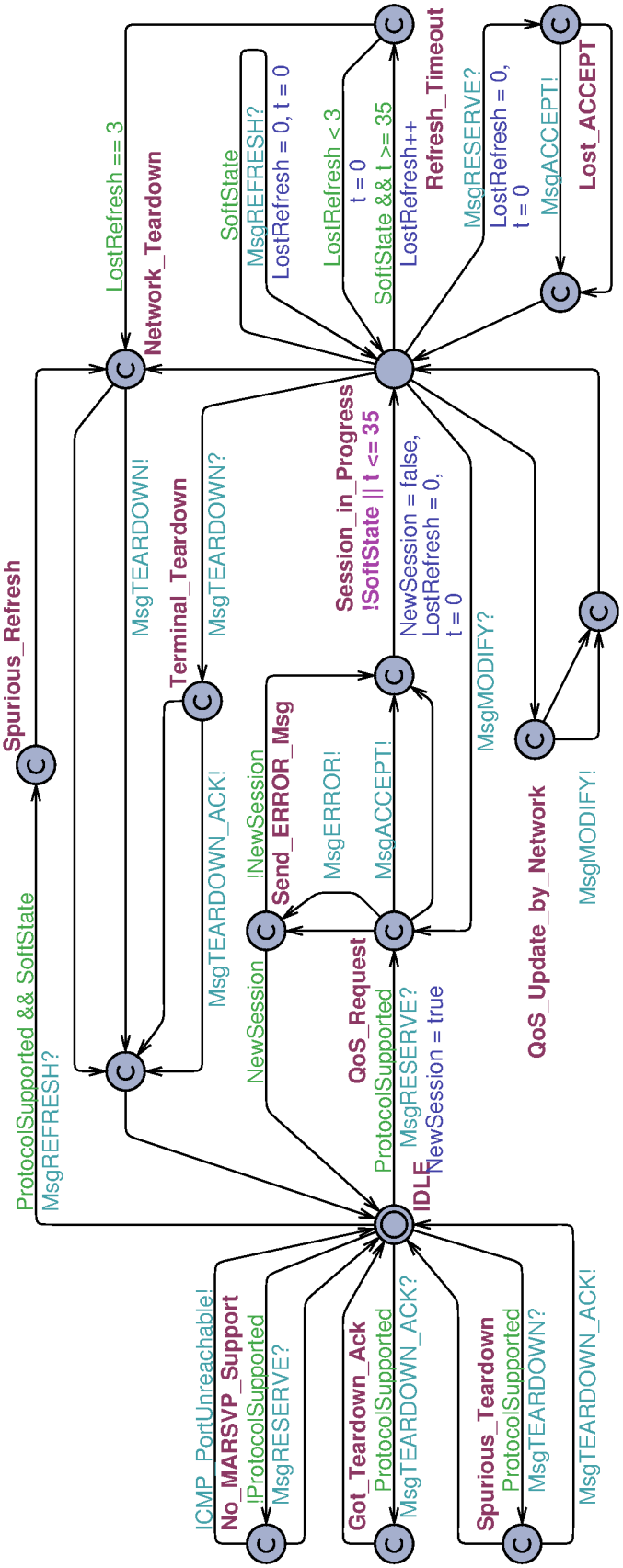


Figure 6.5 State Machine of the Access Router

6.6.1 Properties satisfied by the model

Prior to an implementation of MARSVP into a terminal and an access router, a formal validation of the protocol could prove that the latter works according to specifications. Due to the presence of wireless links, the loss of a message could lead to a desynchronization and eventually to a general failure.

Three kinds of properties are verified in our model:

- **Reachability properties**: these are used to validate the basic behavior of the model by performing sanity checks;
- **Safety properties**: these prove that the model always performs according to specifications;
- **Liveness properties**: these properties verify that given an initial state, some other state will eventually be reached.

The properties we have verified are listed below, followed by a brief description of each one.

1. E<> Application.QoS_Updated_by_Network
2. E<> Application.QoS_Update_Accepted
3. E<> Application.QoS_Update_Rejected
4. E<> Application.Cannot_get_QoS and Terminal.Retry == 3
5. E<> Application.Cannot_get_QoS and Terminal.Retry < 3
6. E<> AccessRouter.Lost_ACCEPT
7. E<> AccessRouter.Got_Teardown_Ack
8. E<> AccessRouter.Spurious_Teardown
9. E<> AccessRouter.Spurious_Refresh
10. A<> Application.IDLE
11. A[] not deadlock
12. E[] AccessRouter.LostRefresh == 3 imply AccessRouter.Network_Teardown
13. A[] Application.QoS_Update_Rejected imply Terminal.NewSession == false
14. A[] Terminal.Send_REFRESH imply SoftState == true
15. A[] Terminal.No_MARSVP_Support imply AccessRouter.ProtocolSupported == false
16. A[] Application.Cannot_get_QoS imply Terminal.NewSession == true
17. Terminal.No_MARSVP_Support → Application.Cannot_get_QoS
18. Terminal.Got_TEARDOWN_Message → Application.Session_Closed_by_Network

19. Terminal.Got_an_ERROR_Message and Terminal.NewSession → Application.Cannot_get_QoS
20. Terminal.Got_an_ERROR_Message and !Terminal.NewSession → Application.QoS_Update_Rejected

Properties #1 to #9 are reachability properties. Property #1 ensures that the network can successfully change the QoS parameters of the Application. Properties #2 and #3 verify that it is possible that a QoS update request be respectively accepted or rejected. Properties #4 and #5 respectively verify that an initial QoS reservation can fail because the `Terminal` didn't get a positive `ACCEPT` message after three attempts or received an `ERROR` message. Property #6 shows that it is possible for the `Access Router` to receive a `RESERVE` message for a resource reservation that has already been accepted (this implies that the `ACCEPT` message was lost in transit). Property #7 ensures that it is possible for the `Access Router` to receive an acknowledgement to a `TEARDOWN` message. Properties #8 and #9 verify that it is possible for an `Access Router` to respectively receive `TEARDOWN` or `REFRESH` messages for sessions that have been previously terminated.

Properties #10 to #16 are safety properties. Properties #10 and #11 respectively verify that the model is free of livelocks and deadlocks. Property #12 verifies that after three lost `REFRESH` messages, the `Access Router` will proceed with a network-initiated teardown of the session [9]. Property #13 ensures that a QoS update request can only be rejected if the session is already established at the application level. Property #14 verifies that a `REFRESH` message can only be sent if we use the soft-state variant of MARSVP. Property #15 ensures that the `Terminal` can get an Internet Control Message Protocol (ICMP) error message only if MARSVP isn't supported by the `Access Router`. Property #16 makes sure that the `Application` will receive a "Cannot Get QoS" error message only if it was trying to create a new session.

Properties #17 to #20 are liveness properties. Property #17 ensures that it is not possible for the `Application` to get any QoS if MARSVP isn't supported. Property #18 ensures that when the `Terminal` receives a `TEARDOWN` message, this forces the `Application` to close its session. Properties #19 and #20 respectively verify that the reception of an `ERROR` message by the `Terminal` causes either an initial resource reservation or an update procedure to fail.

9. Of course, this property can only verified for MARSVP's soft state variant.

6.7 Conclusion

In this paper we validated an access independent resource reservation protocol that enables applications not supported by the network to setup QoS for their needs. Our solution also addresses the problem of dynamically mapping the EPC's QoS classes onto the L2/L3 QoS classes of the underneath transport network. Finally, MARSVP facilitates service convergence by defining a single mechanism to manage QoS, whatever the current access technology being used.

In order to validate the behavior of our solution, we implemented the state machines of the access router, terminal and its application into an Uppaal model. We carefully duplicated all signals that would be sent over the wireless interface so that we can simulate packet losses. Then, we performed a formal validation of the model which demonstrated that the solution always behaves according to specifications.

Our next steps are to fully specify the encoding of QoS mapping informations of several accesses into the MARSVP protocol. Then, MARSVP will be implemented into a terminal and another computer that will act as an access router.

CHAPITRE 7

DISCUSSION GÉNÉRALE

Le présent chapitre a pour but de discuter des propositions faites dans cette thèse en regard des articles que l'on retrouve dans la littérature scientifique au moment où la thèse fut déposée. En effet, la majeure partie de la rédaction des deux premiers articles fut entreprise alors que la version 8 de l'architecture 3GPP était courante et que les travaux commençaient à peine pour la version 9. Il était donc nécessaire de vérifier si les propositions faites sont toujours compatibles avec la version 11 actuelle (en date de septembre 2012).

7.1 Partage d'infrastructures

Le sujet du partage d'infrastructures physiques, tel qu'abordé dans cette thèse, a peu été étudié dans la littérature. En effet, ce sujet a surtout été traité dans le cadre du réseau d'accès, sous forme d'optimisation de l'allocation des ressources radio, de réduction de coûts, de relèves verticales avec minimum d'impacts pour la QoS des flots du terminal, etc.

Cependant, une proposition d'architecture de NGN basée sur la virtualisation est présentée dans [37, 44]. Nous l'avons survolée à la section 2.5.6. Dans cette proposition, les infrastructures physiques sont partagées entre des MVNOs grâce à une entité centralisée connue sous le nom de Network Configuration Platform (NCP).

Le NCP est conceptuellement très semblable au Network Policy Function (NPF) qui fut proposé dans les deux premiers articles (chapitres 4 et 5). En effet, le NCP permet le partage des ressources entre les MVNOs avec une granularité grossière. *A fortiori*, le but premier du NCP (et du NPF) est de cacher les détails des infrastructures physiques sous-jacentes, de telle sorte qu'un plan de contrôle unifié est présenté aux Virtual Network Controllers (VNCs). Enfin, tout comme pour le NPF, le NCP permet de séparer l'application de politiques relatives aux usagers et aux services de celles qui s'appliquent au réseau.

Dans cette thèse, il est important de rappeler que l'architecture PCC est une architecture *logique* déployée au-dessus d'un réseau IP. [1] S'ils étaient rapportés dans l'architecture de [44], nos changements seraient concentrés dans le plan du Virtual Mobile Network (VMN).

La liste de requis élaborée dans [44] ne propose aucune modification à l'actuelle architecture PCC pour mieux supporter le partage d'infrastructures. Par contre, l'un des requis

1. 3GPP ne préconise aucune technologie aux niveaux 1 et 2 dans le réseau cœur. Le seul requis est que le réseau cœur doit pouvoir transporter des paquets IP et supporter DiffServ.

est d'étendre la portée du partage d'infrastructures de façon à partager dynamiquement les ressources entre les MVNOs. Il s'agit très précisément du point que nous avons développé dans les deux premiers articles (chapitres 4 et 5).

Comparativement à [44], notre solution a l'avantage d'être une *évolution* de l'architecture PCC, ce qui peut en faciliter l'acceptation par l'industrie. Au contraire, le NCP nécessite le développement de nouveaux protocoles pour les interfaces reliant les différents plans. Aussi, la virtualisation complique la détection de violations de SLA à cause de l'isolation des plans [44].

7.1.1 Interfaces de PCC

Il est important de revenir sur le second article dans lequel on procède à une modification au niveau des interfaces du NPF. En effet, la modification permet de réconcilier les deux variantes de l'architecture PCC (GTP et PMIPv6) et d'offrir un plan progressif de migration vers l'architecture proposée dans le second article. Les relations suivantes représentent le contenu des interfaces Gx et Gxc de 3GPP pour les variantes PMIPv6 et GTP :

$$
\begin{aligned}
Gxc_{PMIPv6} &= \{\text{QoS, } \textit{bearer binding}, \textit{ event reporting}\} \\
Gx_{PMIPv6} &= Gxc_{PMIPv6} \setminus \{\textit{bearer binding}\} \cup \{\textit{charging information}, \textit{ gating control}\} \\
Gx_{GTP} &= Gx_{PMIPv6} \cup Gxc_{PMIPv6} \\
Gxc_{GTP} &= \varnothing
\end{aligned}
$$

Il faut mentionner que les interfaces décrites précédemment dérivent d'une spécification unique. Les interfaces qui relient le NPF aux PCRF, S-GW et P-GW sont des extensions de Gx_{PMIPv6} et Gxc_{PMIPv6} incluant le support des *Infrastructure Services* (IS) et des Border Gateway Functions (BGFs). [2] Le Tableau 7.1 illustre les relations entre les variantes de l'architecture 3GPP et de celles présentées dans les deux premiers articles (chapitres 4 et 5).

Tableau 7.1 Comparaison des interfaces pour 3GPP et nos deux propositions

Interface	3GPP R8	Article #1	Article #2
Applic. ↔ PCRF	Rx	Rx + IS	Rx + IS
PCRF ↔ P-GW	Gx	Gx - QoS + BGF	N/A
PCRF ↔ S-GW	Gxc	N/A	N/A
PCRF ↔ NPF	N/A	Gxc + IS	Gx + IS
NPF ↔ P-GW	N/A	Gxc + IS	Gx + IS
NPF ↔ S-GW	N/A	Gxc + IS	Gxc + IS

2. Notez que BGF ⊆ IS et que ces fonctions sont implémentées dans le P-GW du côté opérateur.

7.1.2 Évolution de l'architecture PCC des réseaux 3GPP

L'architecture PCC du EPS a évolué depuis ses débuts dans la version 8 jusqu'à nos jours (version 11). La Figure 2.5 à la page 19 présente l'architecture PCC pour la version 8 tandis que la Figure 7.1 illustre la version 11 de cette dernière. On peut compter deux changements significatifs comparativement à la version 8 :

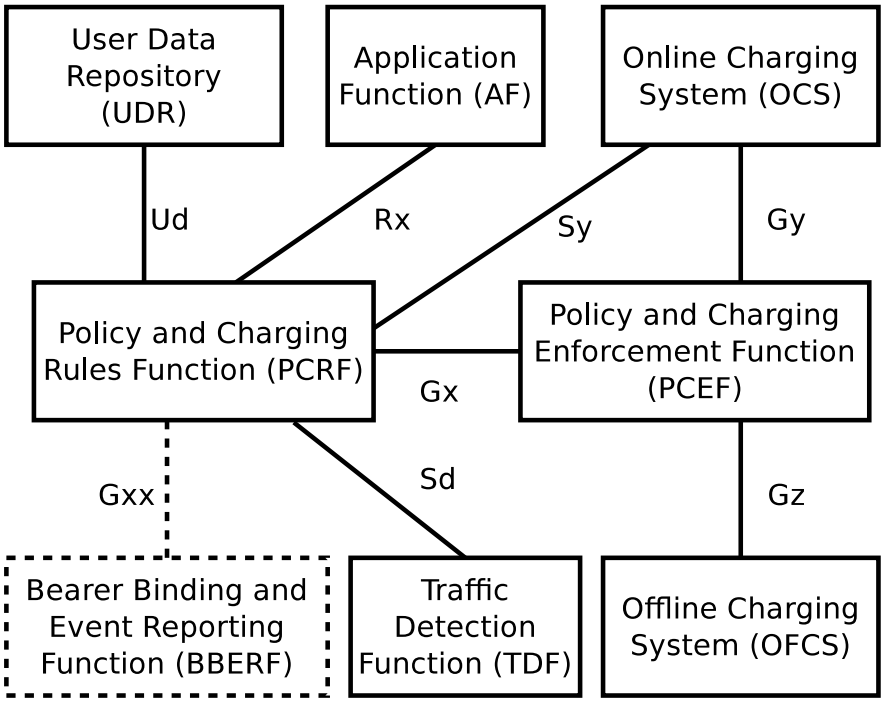
1. la version 11 de l'architecture PCC supporte le concept d'Application Detection and Control (ADC) qui permet d'informer le PCRF lorsqu'un flot de données est initié ou terminé et d'appliquer les politiques de réseau appropriées. Le support ADC peut être incorporé dans un PCEF amélioré ou un nouveau nœud logique nommé Traffic Detection Function (TDF). Le TDF peut être configuré statiquement pour détecter un certain nombre d'applications et en informer le PCRF ou être configuré dynamiquement selon le profil de l'usager ;
2. l'autre changement qui caractérise maintenant l'architecture PCC est l'introduction de l'architecture User Data Convergence (UDC) en remplacement du SPR. En effet, l'architecture UDC [4] est caractérisée par une séparation de la logique des applications des données des usagers. Ainsi, il est possible de regrouper toutes les informations concernant un usager dans une seule base de données [3] qui est accessible de toutes les applications auxquelles l'usager a souscrit. À titre d'exemple, il devient possible pour une application basée sur la géolocalisation d'être informée lorsqu'arrive un événement donné puisque toutes les informations de l'usager sont regroupées et accessibles.

Lorsque les changements introduits depuis la version 8 de l'architecture PCC sont rapportés à nos propositions des chapitres 4 et 5, l'analyse de ces modifications conclut que :

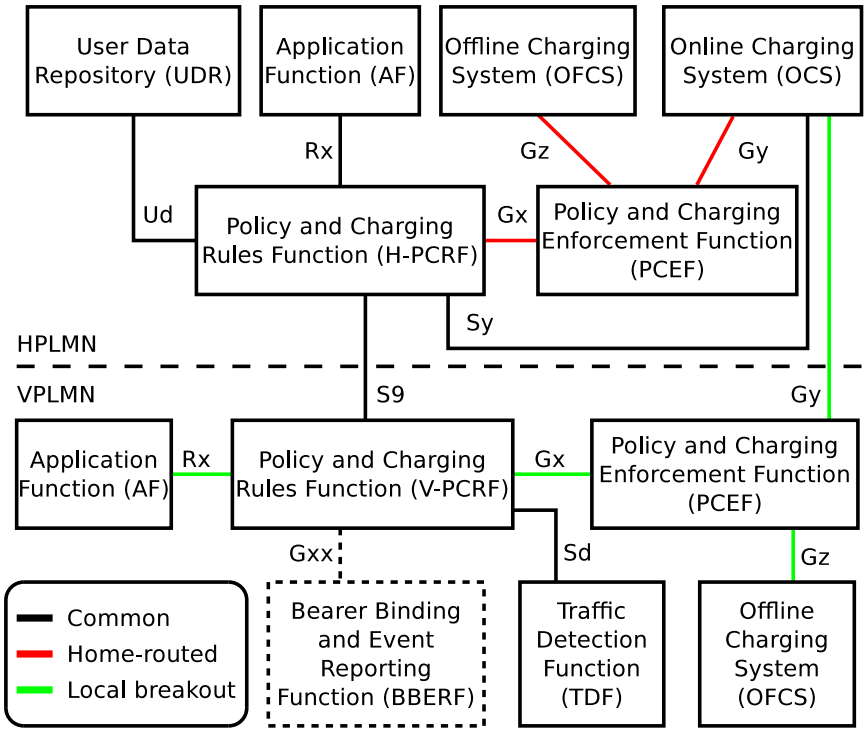
- le support de l'architecture UDC serait confiné à un NSP et ne serait donc pas perceptible à partir de l'IPAN ;
- le support d'ADC ne nécessite aucun changement si la TDF est intégrée à même le PCEF. Par contre, une interface Sd serait requise entre le NPF et un nœud TDF implémenté hors du PCEF. L'analyse de l'interface Sd révèle que cette dernière est dérivée de Gx, ce qui rend le NPF capable d'extraire de façon transparente les informations requises par le nœud TDF. Dans tous les cas, le PCRF doit agir comme si le TDF était intégré dans le PCEF.

En conclusion, l'évolution de l'architecture PCC de la version 11 n'introduit pas d'incompatibilité qui rendrait caduques nos propositions d'améliorations telles que présentées dans les chapitres 4 et 5.

3. Cette base de données unique porte le nom de User Data Repository (UDR).



(a) Architecture sans itinérance.



(b) Cas d'itinérance *Home-routed* et *local breakout*.

Figure 7.1 Architecture PCC logique avec support UDC pour la version 11

7.2 Discussion sur la QoS multiaccès

Au chapitre 6, MARSVP a adressé un problème qui n'avait pas été auparavant soulevé dans la littérature. Malgré cela, son importance se justifie par la nécessité de permettre l'émergence d'applications non-supportées par le réseau. En conséquence, MARSVP spécifie les classes de services dynamiquement, ce qui est utile pour différencier le traitement appliqué aux usagers ou lors des cas d'itinérance.

Finalement, il faut mentionner que l'évolution de l'architecture PCC à la version 11 n'aura pas d'impact sur le fonctionnement de MARSVP ni sur sa pertinence en tant que mécanisme de support de la QoS. En effet, l'introduction du nœud logique TDF ne peut suppléer à MARSVP puisque celui-ci ne peut traiter que des flots de données connus tandis que MARSVP vise à offrir un support de la QoS pour les applications nouvelles ou inconnues du réseau.

CHAPITRE 8

CONCLUSION ET RECOMMANDATIONS

Ce chapitre permet d'effectuer un retour sur le travail accompli dans cette thèse. Tout d'abord, nous ferons un rappel des principales contributions de la thèse et mettrons en évidence l'originalité de l'approche que nous avons prise. Ensuite, nous discuterons des limitations actuelles de nos travaux. Enfin, nous proposerons des avenues de travaux possibles qui découlent des accomplissements de cette thèse ainsi que de ses limitations.

8.1 Principales contributions et originalité de la thèse

Les principales contributions de cette thèse permettent de mieux supporter la FMC de même que le partage d'infrastructures au niveau du réseau cœur entre plusieurs opérateurs. Les moyens utilisés se présentent sous forme d'améliorations à l'architecture PCC actuelle.

Dans un premier temps, une liste des tendances lourdes au sein de l'industrie des communications mobiles fut dressée. Ces tendances ont en principe une influence directe et significative sur l'architecture PCC recherchée. Les tendances identifiées sont les suivantes :

La FMC qui permet d'offrir soit la meilleure couverture possible ou celle qui est la plus efficiente en termes de coûts. De plus, la FMC permet d'offrir une couverture à des endroits où ce ne serait généralement pas rentable.

Le partage d'infrastructures qui devient rentable car le revenu moyen par usager est demeuré constant depuis longtemps alors que les investissements des opérateurs sont en forte hausse pour maintenir les infrastructures à jour.

La spécialisation des opérateurs est une tendance caractérisée par les moyens entrepris par les opérateurs pour se démarquer de leurs compétiteurs, entre autres par une offre de services adaptée à une clientèle spécifique.

La désagrégation du modèle vertical d'opérateur est causée par la volonté de ceux-ci de concentrer leurs efforts sur les aspects de leurs opérations qui ont la plus grande valeur ajoutée et de sous-traiter plusieurs de leurs opérations courantes.

Dans un second temps, l'analyse des architectures PCC de 3GPP et du RACS de TISPAN a permis d'élaborer une liste de quatre concepts qui constituent une liste de requis que l'architecture doit supporter afin de simplifier le support de la FMC et du partage d'infrastructures.

Ensuite, nous avons défini des rôles d'affaires qui permettent de diviser les tâches entre les intervenants et ainsi réduire le nombre de combinaisons possibles de relations d'affaires.

Les rôles d'affaires sont :

Le Access Network Provider (ANP) qui gère les ressources d'un réseau d'accès.

Le NSP qui gère une banque d'usagers et offre à ces derniers des services avancés.

Le IPAN Provider qui contrôle les ressources du réseau cœur et les partage entre plusieurs NSPs.

Finalement, la liste des tendances de l'industrie, de celle des concepts directeurs, de même que les définitions des rôles d'affaires ont servi d'intrants à la proposition de solution. En effet, nous avons introduit le NPF afin de contrôler le IPAN. Le NPF est responsable de partager les ressources du réseau entre les MVNOs, de leur cacher les détails de l'implémentation du IPAN et de gérer les *Infrastructure Services*.

Dans le second article (chapitre 5), nous avons mis en évidence les nombreuses simplifications potentielles au niveau des nœuds du réseau cœur. Par exemple, dans les cas d'itinérance, nous forçons toujours le trafic d'un terminal à passer par le P-GW du réseau visité. Ainsi, nous pouvons simplifier le S-GW en retirant toutes les fonctions de comptabilisation des ressources de même que la capacité du S-GW à émettre des Charging Data Records (CDRs).

De plus, dans le second article, nous avons procédé à une modification de notre solution, qui semble anodine à première vue, mais qui a comme impact d'uniformiser notre solution aux deux variantes du EPC (basées sur GTP et PMIPv6).

L'originalité de la présente recherche repose sur le fait que l'évolution proposée à l'architecture PCC origine de la flexibilité désirée en ce qui concerne les scénarios de partage d'infrastructures et de support de la FMC. En effet, la séparation des rôles d'affaires permet d'intégrer facilement un nouveau joueur à un réseau existant, tandis que les opérateurs traditionnels seront supportés par la nouvelle architecture en occupant plus d'un rôle. À titre d'exemple, un opérateur possédant *a priori* toutes les infrastructures pourra évoluer vers un scénario de partage de son réseau cœur d'une façon naturelle, sans devoir procéder à des adaptations majeures de son architecture puisque celle-ci est déjà prête à être partagée. *A contrario*, un opérateur désirant se départir d'une partie de ses infrastructures, dans le but de les louer au futur acquéreur, pourra le faire sans procéder à un reconfiguration majeure.

Par ailleurs, le protocole MARSVP qui fut proposé au chapitre 6 apporte une solution aux applications qui ne bénéficient d'aucun support de la QoS de la part du réseau. De plus, MARSVP spécifie les associations entre les QCIs de 3GPP et les classes de services aux niveaux 2 et 3. Enfin. notre protocole permet de respecter la philosophie du EPS pour lequel l'activation du support de la QoS est initié par le réseau.

8.2 Limitations de la thèse

La nature même du sujet de recherche de cette thèse rend difficile toute implémentation qui serait utile afin de valider les solutions proposées dans un environnement réel. Par ailleurs, une implémentation qui ne serait pas faite dans du matériel comparable (en termes de performances) à celui qui est déployé en situation réelle ne pourrait donner qu'un aperçu des impacts de nos propositions sur l'architecture PCC.

Aussi, il est important de rappeler que l'architecture PCC est une architecture *logique*. En effet, l'ajout d'un nœud supplémentaire au sein de l'architecture PCC en dit peu sur l'impact global de notre proposition. Il est courant que les implémentations physiques combinent plusieurs nœuds logiques. Par exemple, la compagnie Airspan propose un EPC complet dans un seul boîtier qui peut supporter jusqu'à 200 000 usagers simultanément.

8.3 Travaux futurs

Les sous-sections suivantes abordent divers sujets ayant un rapport direct avec la thèse ou constituent des sujets à considérer dans le futur en raison de leurs interactions avec le EPC.

8.3.1 Infrastructure Services

Dans les chapitres 4 et 5, le concept d'*Infrastructure Services* a été survolé mais peu élaboré. En effet, leur raison d'être repose sur la valeur ajoutée à l'IPAN et à la possibilité d'introduire des services qui soient partageables entre les MVNOs. De plus, le rattachement de ces services aux infrastructures physiques constitue une forte justification de leur existence. [1]

Dans un autre ordre d'idée, la désintégration du modèle vertical d'opérateur peut en amener certains à se départir de services essentiels mais à faible valeur ajoutée tels qu'un antivirus au niveau réseau ou une *cache* pour fichiers multimédias. [2] Enfin, les *Infrastructure Services* peuvent offrir des filtres de trafic indésirable en raison des coûts de l'interface radio.

D'autres travaux seront nécessaires afin de définir des mécanismes d'annonce et d'appel des *Infrastructure Services*. Le but est de standardiser un certain nombre de fonctions de base, par exemple comment un NSP demande un transcodage en temps-réel d'un fichier multimédia, ou comment une application de distribution de contenu multimédia demande de mettre en cache un fichier qui sera très souvent relu par ses clients, etc.

1. De bons exemples seraient des services d'alertes diffusées à grande échelle, d'appel d'urgence ou d'interception d'appel pour les forces de l'ordre, de *cache* pour des fichiers multimédias, etc.

2. Ce dernier service est très utile pour décharger les serveurs d'un NSP tout en diminuant la consommation de bande passante dans le EPC.

8.3.2 Interface Web usager

L'industrie des communications mobiles considère que les services offerts par le EPS seront des services définis uniquement par les opérateurs cellulaires. Toutefois, cela pourrait se révéler faux en raison de la force de l'Internet et son omniprésence.

En conséquence, l'interface Rx qui relie le AF au PCRF dans l'EPC et qui est basée sur le protocole Diameter pourrait subir des pressions dans le but de la faire évoluer vers un protocole beaucoup plus « *Web-friendly* », basé sur une représentation textuelle. C'est le cas pour le très populaire HyperText Transfer Protocol (HTTP), le Simple Mail Transfer Protocol (SMTP) de même que pour le Session Initiation Protocol (SIP).

8.3.3 Near Field Communications et Fournisseur d'identité

L'évolution des modes de paiement a récemment vu apparaître les Near Field Communications (NFCs). Les NFCs sont basées sur la technologie de Radio-Frequency Identification (RFID) appliquée des appareils cellulaires. Elles permettent des communications à très courtes distances (de l'ordre de quelques centimètres) et sont conçues pour l'identification et l'authentification des usagers dans le but d'effectuer des paiements ou des échanges sécurisés.

Parallèlement, le rôle de NSP tel que présenté dans les chapitres 4 et 5 comprend à la fois celui de fournisseur de services et celui d'authentification d'usager. Il serait souhaitable que ces rôles soient séparés de telle sorte que des compagnies spécialisées en crédit et en authentification puissent contribuer à évoluer les modes de paiement sans-fil.

8.3.4 Machine-to-Machine (M2M) Communications

Des manufacturiers d'équipements de communications mobiles prévoient que vers 2020 le nombre d'appareils de toutes sortes qui seront connectés à Internet dépassera le seuil des cinquante milliards. Chaque appareil sera doté de son propre Subscriber Identity Module (SIM) et aura à sa disposition toutes les technologies câblées ou sans-fil pour communiquer.

Les communications M2M accapareront une portion importante du trafic sur Internet et comprendront entres autres des réseaux de senseurs, des distributrices de boissons gazeuses, des robots d'usine, des compteurs électriques intelligents, etc.

Il est difficile de prévoir quelles sont les contraintes spécifiques que les communications M2M du futur imposeront au EPC car les utilisations actuelles sont surtout de nature transactionnelle et que celles-ci n'ont que des besoins en QoS très modestes. Toutefois, leur seul nombre peut poser un défi de gestion des sessions au niveau de l'architecture PCC.

8.3.5 Implémentation *Cloud* des nœuds du EPC

La tendance actuelle qui consiste à virtualiser des serveurs de bases de données et d'autres applications en général n'échappe pas au monde des communications mobiles. En effet, nous avons survolé à la section 2.5.6 une proposition d'architecture [44] qui compte virtualiser l'ensemble du EPC et ainsi transformer tous les NSPs en MVNOs.

Le « *cloud computing* » offre des possibilités uniques à l'industrie des communications mobiles. En effet, il serait possible de construire un seul nœud logique à partir de plusieurs machines virtuelles qui travaillent en collaboration. Ainsi, on pourrait ajuster la capacité du nœud en fonction de la charge de travail à supporter. De plus, la redondance inhérente à la collection des machines virtuelles permettrait d'augmenter la fiabilité générale du nœud logique. Par ailleurs, une collection de machines physiques de série coûterait vraisemblablement moins cher que du matériel dédié à cette tâche. Finalement, le véritable défi de cette approche consiste à cacher le fait qu'un nœud logique de l'architecture PCC est en réalité composé de multiples machines virtuelles qui travaillent en parallèle.

8.3.6 Simulation des performances de l'architecture proposée

Les articles portant sur le partage d'infrastructures (chapitres 4 et 5) ne présentent aucun résultat de simulation des impacts qu'auraient l'introduction du NPF dans un EPC partagé entre deux MVNOs. En effet, ces travaux furent réalisés dans le cadre d'une architecture PCC *logique*.

Bien que les traitements effectués par le NPF soient uniquement liés à la signalisation et non au transport des données, il importe de mesurer les délais supplémentaires dus à sa présence. En conséquence, la conception d'un modèle d'EPC pour un simulateur tel qu'Opnet pourrait permettre d'évaluer les impacts du NPF sur le temps total requis pour établir un nouveau flot de données. Il faudra procéder à cette évaluation pour plusieurs technologies d'accès de même que pour les variantes de quelques unes d'entre elles [3] afin de déterminer si le NPF produit un impact négligeable ou significatif.

3. Par exemple, LTE supporte les variantes Time-Division Duplexing (TDD) et Frequency-Division Duplexing (FDD).

RÉFÉRENCES

[1] 3GPP. *TS 22.951 : Service Aspects and Requirements for Network Sharing.*

[2] 3GPP. *TS 23.203 : Policy and Charging Control Architecture.*

[3] 3GPP. *TS 23.251 : Network Sharing ; Architecture and Functional Description.*

[4] 3GPP. *TS 23.335 : User Data Convergence (UDC).*

[5] 3GPP. *TS 23.401 : General Packet Radio Service Enhancements for E-UTRAN Access.*

[6] 3GPP. *TS 23.402 : Architecture Enhancements for non-3GPP Accesses.*

[7] 3GPP. *TS 29.276 : Optimized Handover Procedures and Protocols between E-UTRAN Access and cdma2000 HRPD Access.*

[8] 3GPP2. *X.S0057-0 : E-UTRAN - eHRPD Connectivity and Interworking : Core Network Aspects.*

[9] 4WARD. *D-3.1.1 - Virtualization Approach : Concepts*, 2009.

[10] L. Anderson and G. Swallow. *RFC3468 : The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols.* IETF, February 2003. Status : INFORMATIONAL.

[11] C. Beckman and G. Smith. Shared Networks : Making Wireless Communications Affordable. *IEEE Wireless Communications*, 12(2) :78–85, April 2005.

[12] I. L. Bedhiaf, O. Cherkaoui, and G. Pujolle. Third-Generation Virtualized Architecture for the MVNO Context. *Annales des Télécommunications/Annals of Telecommunications*, 64(5–6) :339–347, 2009.

[13] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin. *RFC2205 : Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification.* IETF, September 1997. Status : STANDARDS TRACK.

[14] V. Daniel Philip, Y. Gourhant, and D. Zeghlache. Preliminary Analysis of 4G-LTE Mobile Network Sharing for Improving Resiliency and Operator Differentiation. In *Communications in Computer and Information Science*, volume 171, pages 73–93, 2011.

[15] H. Ekström. QoS Control in the 3GPP Evolved Packet System. *IEEE Communications Magazine*, 47(2) :76–83, February 2009.

[16] K. Evensen, D. Kaspar, P. Engelstad, A.F. Hansen, C. Griwodz, and P. Halvorsen. A network-layer proxy for bandwidth aggregation and reduction of ip packet reordering. In *Local Computer Networks, 2009. LCN 2009. IEEE 34th Conference on*, pages 585–592, October 2009.

[17] T. K. Forde, I. Macaluso, and L. E. Doyle. Exclusive sharing virtualization of the cellular network. In *2011 IEEE International Symposium on Dynamic Spectrum Access Networks, DySPAN 2011*, pages 337–348, May 2011.

[18] J.C. Francis. Techno-economic analysis of the open broadband access network wholesale business case. In *Mobile and Wireless Communications Summit, 2007. 16th IST*, pages 1–5, July 2007.

[19] T. Frisanco. Strategic and economic benefits of regionalization, centralization, and outsourcing of mobile network operations processes. In *5th International Conference on Wireless and Mobile Communications, ICWMC 2009*, pages 285–290, 2009.

[20] T. Frisanco, P. Tafertshofer, P. Lurin, and R. Ang. Infrastructure Sharing and Shared Operations for Mobile Network Operators from a Deployment and Operations View. In *NOMS 2008 - IEEE/IFIP Network Operations and Management Symposium : Pervasive Management for Ubiquitous Networks and Services*, pages 129–136, 2008.

[21] GENI. *Global Environment for Network Innovations (GENI) System Overview*, 2008.

[22] H. Haffajee and H. A. Chan. Low-cost qos-enabled wireless network with interworked wlan and wimax. In *Auswireless 2006 Conference*, 2006.

[23] R. Hancock, G. Karagiannis, J. Loughney, and S. Van den Bosch. *Next Steps in Signaling (NSIS) : Framework*. IETF, June 2005. Status : INFORMATIONAL.

[24] M. Hoffmann and M. Staufer. Network Virtualization for Future Mobile Networks. In *IEEE International Conference on Communications (ICC) Workshops*, 2011.

[25] IEEE. *IEEE Standard for Local and Metropolitan Area Networks - Medium Access Control (MAC) Bridges*, 2004.

[26] IEEE. *IEEE Standard for Local and Metropolitan Area Networks - Virtual Bridged Local Area Networks*, 2005.

[27] IEEE. *IEEE Standard for Local and Metropolitan Area Networks - Virtual Bridged Local Area Networks – Amendment 4 : Provider Bridges*, 2005.

[28] IEEE. *IEEE Standard for Local and Metropolitan Area Networks - Part 11 : Wireless LAN Medium Access Control and Physical Layer Specifications*, 2007.

[29] IEEE. *IEEE Standard for Local and Metropolitan Area Networks - Virtual Bridged Local Area Networks – Amendment 7 : Provider Backbone Bridges*, 2008.

[30] IEEE. *IEEE Standard for Local and Metropolitan Area Networks - Virtual Bridged Local Area Networks – Amendment 7 : Provider Backbone Bridges Traffic Engineering*, 2009.

[31] IEEE. *IEEE Standard for Local and Metropolitan Area Networks - Part 11 : Wireless LAN Medium Access Control and Physical Layer Specifications - Amendment 9 : Interworking with External Networks*, 2011.

[32] IETF. *RFC2113 : IP Router Alert Option*, February 1997. Status : PROPOSED STANDARD.

[33] IETF. *RFC2711 : IPv6 Router Alert Option*, October 1999. Status : PROPOSED STANDARD.

[34] IETF. *RFC 3945 : Generalized MultiProtocol Label Switching (GMPLS) Architecture*, October 2004. Status : PROPOSED STANDARD.

[35] ITU-T. *Recommendation Y.2001 - Next Generation Networks : Frameworks and Functional Architecture Overview*, December 2004.

[36] D. Kataria and D. Logothetis. Fixed mobile convergence : network architecture, services, terminals, and traffic management. In *Personal, Indoor and Mobile Radio Communications, 2005. PIMRC 2005. IEEE 16th International Symposium on*, volume 4, pages 2289–2300, September 2005.

[37] A. Khan, D. Jurca, K. Kozu, W. Kellerer, and M Yabusaki. The Reconfigurable Mobile Network. In *IEEE International Conference on Communications (ICC)*, June 2011.

[38] S. Krishnan, L. Marchand, and G. N. Cassel. An IETF-based Evolved Packet System beyond the 3GPP Release 8. In *CTIA - The Wireless Association*, 2008.

[39] T. Li and L. Bai. Model of wireless telecommunications network infrastructure sharing and benefit-cost analysis. *International Conference on Information Management, Innovation Management and Industrial Engineering*, 2 :102–105, 2011.

[40] S. Mangold, S. Choi, G. R. Hiertz, O. Klein, and B. Walke. Analysis of IEEE 802.11e for QoS Support in Wireless LANs. *IEEE Wireless Communications*, 10(6), December 2003.

[41] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner. OpenFlow : Enabling Innovation in Campus Networks. *ACM SIGCOMM Computer Communication Review*, 38(2), March 2008.

[42] D.-E. Meddour, T. Rasheed, and Y. Gourhant. On the role of infrastructure sharing for mobile network operators in emerging markets. *Computer Networks*, 55(7) :1576 – 1591, 2011.

[43] P. Mell and T. Grance. *The NIST Definition of Cloud Computing (Version 15)*. National Institute of Standards and Technology, Information Technology Laboratory, USA, October 2009.

[44] Network Sharing in the Next Mobile Network : TCO Reduction, Management Flexibility, and Operational Independence. A. khan and w. kellerer and k. kozu and m. yabusaki. *IEEE Communications Magazine*, 49(10) :134–142, October 2011.

[45] M. Olsson, S. Sultana, S. Rommer, L. Frid, and C. Mulligan. *SAE and the Evolved Packet Core*. Academic Press, 2009.

[46] S. Ouellette, L. Marchand, and S. Pierre. A Potential Evolution of the Policy and Charging Control/QoS Architecture for the 3GPP IETF-based Evolved Packet Core. *IEEE Communications Magazine*, 49(5) :231–239, May 2011.

[47] J-J. Pastor Balbás, S. Rommer, and J. Stenfelt. Policy and Charging Control in the Evolved Packet System. *IEEE Communications Magazine*, 47(2) :68–74, February 2009.

[48] E. Rosen, A. Viswanathan, and R. Callon. *RFC 3031 : MultiProtocol Label Switching (MPLS) Architecture*. IETF, January 2001. Status : STANDARDS TRACK.

[49] R. Sherwood, M. Chan, G. Gibb, N. Handigol, T.-Y. Huang, P. Kazemian, M. Kobayashi, D. Underhill, K.-K. Yap, G. Appenzeller, and Nick McKeown. Carving Research Slices out of your Production Network with OpenFlow. *ACM SIGCOMM Computer Communication Review*, 40(1), January 2010.

[50] G. Smith and C. Beckman. Shared Networks : More than Making Wireless Communications Affordable. In *IEEE 61st Vehicular Technology Conference*, volume 5, pages 2984–2988, 2009.

[51] TISPAN. *ETSI Standard 282 001 : Telecommunications and Internet-converged Services and Protocols for Advanced Networks (TISPAN) : Next Generation Network (NGN) Functional Architecture*.

[52] TISPAN. *ETSI Standard 282 003 : Telecommunications and Internet-converged Services and Protocols for Advanced Networks (TISPAN) : Resource and Admission Control Subsystem (RACS) : Functional Architecture*.

[53] W. Webb, editor. *Wireless Communications - the Future*. Wiley, 2007.

[54] M. Yabusaki and Y. Okumura. Next Mobile Network Architecture. In *IEEE International Conference on Communications (ICC)*, May 2010.